DOCUMENT RESUME

ED 355 238 TM 019 540

AUTHOR	Levitan, Sar A.
TITLE	Evaluation of Federal Social Programs: An Uncertain Impact. Occasional Paper 1992-2.
INSTITUTION	George Washington Univ., Washington, D.C. Center for Social Policy Studies.
SPONS AGENCY	Ford Foundation, New York, N.Y.
PUB DATE	Jun 92
NOTE	70p.
AVAILABLE FROM	Public Interest Publications, 3030 Clarendon Boulevard, Suite 200, Arlington, VA 22201, or P.O. Box 229, Arlington, VA 22210.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Decision Making; *Evaluation Methods; Evaluation Utilization; *Experiments; *Federal Programs; *Program Evaluation; Public Policy; *Qualitative Research; Research Methodology; Simulation; *Social Services
IDENTIFIERS	Department of Health and Human Services; Department of Labor; General Accounting Office

ABSTRACT

This paper explores the impact that the evaluation industry has had on the development and implementation of social policy and programs, primarily as carried out by the U.S. Departments of Labor and Health and Human Services. In addition, major tools evaluators have developed and used, and the institutional arrangements through which they have worked are reviewed. Three principal approaches that have emerged in the evaluation industry funded by executive and congressional agencies and by the grant and contract establishment are microsimulation, experimental and quasi-experimentation, and qualitative studies. With regard to experimentation and quasi-experimentation, a case for randomization, selection modeling, the New Jersey income support experiment, selecting a probabilistic sample, the elusiveness of the target population, attrition, the need to generalize to untested treatments, biases arising from the limited duration of experiments, feedback effects, Hawthorne effects, and the elusiveness of consensus are considered. Advocates of each method may claim that their method of estimating program impact offers policymakers better insights into the effects of policy decisions than those of alternative methods. Analysts have been unable to agree about the impact of social policy evaluation. Current practices, including those used by the General Accounting Office, are generally not adequate as guides to making choices among social policy alternatives. Improved results will be obtained only by making more effective use of various methodologies and by achieving a better balance in funding them. There is cause for optimism that program evaluation will justify its continued funding. (SLD)

Evaluation of Federal Social Programs: An Uncertain Impact

ED355238

TM019540

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

SAR A. LEVITAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

Evaluation of Federal Social Programs: An Uncertain Impact

by
Sar A. Levitan

OCCASIONAL PAPER 1992-2
JUNE 1992

Center for Social Policy Studies
The George Washington University
1717 K Street, NW, Suite 1200
Washington, DC 20006
(202) 833-2530

Contents

Preface iv
Acknowledgements vi
A New Industry 2
The Planning-Programming-Budgeting System 2
Expanding Evaluation Research 4
Microsimulation 5
Experimentation and Quasi-Experimentation 12
The Case for Randomization 14
Selection Modeling 15
From Sunday's Sermon to Monday's Work 19
The New Jersey Income Support Experiment 20
Selecting a Probabilistic Sample 23
The Black Box 25
Elusiveness of the Target Population 25
Attrition 27
The Need to Generalize to Untested Treatments 28
Biases Arising From the Limited
Duration of Experiments 29
Feedback Effects 29
Hawthorne Effects 29
Elusiveness of Consensus 30
Income Maintenance 31
Criminal Justice 33
The Claims of Experimentalists 34
Qualitative Evaluation 37
The Evaluation Industry 42
Executive Agencies 43
Congressional Agencies 44
The Grant and Contract Establishment 50
Toward a More Balanced Evaluation Agenda 52
The Uncertain Returns 59

Evaluation of Federal Social Programs: An Uncertain Impact was prepared under an ongoing grant from the Ford Foundation to the Center for Social Policy Studies of The George Washington University. In line with the Foundation's practice, responsibility for the contents of this study was left with the author.

Preface

Aided by technological progress, social science researchers have achieved during the past three decades significant advances in methodologies and techniques for evaluating federal social programs. Cuts in the funding of evaluation research during the 1980s may have impaired progress. Whether more generous funding would have enabled analysts to come up with important breakthroughs remains speculative.

Analysts have designed several methods for evaluating social programs. Microsimulation is capable of providing policymakers with ball-park cost estimates of proposed initiatives or changes in existing programs, but it is not designed to predict or measure program outcomes. Experimentalists use random assignment as the central or critical element to estimate the impact of a program. Quasi-experimentalists employ other statistical approaches to achieve the same results. All efforts to quantify net impact have inherent limitations, because the estimates are subject to differing ranges of uncertainty and the applicability of estimates based on samples to a national program remains problematic.

Qualitative evaluations attempt to provide insights about the reactions of program participants, the problems encountered by the program, and what might be done to improve program performance. Qualitative evaluation is not designed to estimate the average impact of a program on participants.

Evaluators should — but frequently do not — use a combination of methods in pursuing their tasks. They have not devoted adequate attention to the complexities resulting from multiple evaluation objectives. Evaluators have often ignored the fact that different objectives can best be accomplished by alternative methodologies. New developments offer promise for nonexperimental approaches to the evaluation of social programs.

Institutional factors have set in that may have impeded potential progress. Experimentalists have succeeded in obtaining the lion's share of federal executive agencies' evaluation funds by claiming that other methods of evaluating federal social programs are inferior. Theory is on their side, but they seem to have failed to deliver on their promises.

Experiments dealing with human beings cannot be manipulated as natural scientists can do in a laboratory.

The cost of social experiments also should not be ignored. Considering the sometimes doubtful findings of evaluations and their scant utilization by policymakers, funders might aim for a more balanced distribution of support among evaluation methodologies. The transfer of funds earmarked for evaluation to the development of databases or services to the clients of social programs also deserves consideration.

Given the growth of the evaluation industry and the diverse methodologies developed over the past three decades, it was necessary to selectively limit the subjects to be covered. This paper focuses on the impact the evaluation industry has had on the development and implementation of social policy and programs primarily as carried out by the U.S. Departments of Labor and Health and Human Services. The paper also reviews the major tools that evaluators have developed and utilized, and the institutional arrangements through which they have practiced their trade.

Acknowledgements

I am indebted to Lee Bawden, Seymour Brandwein, Eleanor Chelimsky, Frank Gallo, James Heckman, Garth Mangum, Demetra Nightingale, and Peter Rossi for insightful and helpful comments, and to Keith Hurt and Lawrence Wohl for their contributions to the initial draft of this paper. Their help improved the contents, but they are absolved from any responsibility for the final product. I am also grateful to Miriam Washington for shepherding this paper through several versions and providing editorial clarifications.

Evaluation of Federal Social Programs: An Uncertain Impact

The vast expansion of the American welfare system, during the 1960s and 1970s, was accompanied by sustained attempts to assess whether the efforts achieved the intended goals. Rising federal outlays were sufficient reason for generating interest in evaluating social programs. President Johnson's Great Society initiatives supplied additional pressures for assessing new social programs. Ironically, Congress and the public at large indicated little interest in appraising the value of pork barrel legislation or universal programs that benefited the rich and the poor, while programs designed to help the poor attracted intense congressional and media scrutiny.

Congress mandated that the agencies responsible for implementing the new initiatives periodically assess the programs, and in some cases Congress even earmarked a portion of the appropriated funds for research and evaluation. Congress and the executive branch anticipated that evaluation of program effectiveness would also serve as a basis for the allocation of federal dollars. Starting in the 1960s, the clamor for evaluation spawned a new industry, employing today tens of thousands doing multi-million dollars of business each year.

Whether the evaluators lived up to the expectations of policymakers, and the impact of evaluation on policy, are continuing subjects of debate. The obstacles impeding evaluations have been formidable. Policymakers and program administrators frequently have displayed little interest in the evaluation findings because other considerations influenced their actions. The need for evaluation invariably exceeded available funds, and the distribution was highly concentrated in a few projects, leaving crumbs for other potentially innovative efforts. The slow pace of evaluations has also played a significant role. Not infrequently, by the time evaluators delivered their products, the interest of policymakers in the subjects under investigation waned or they have acted without the benefit of the evaluation findings.

A New Industry

Modern evaluations of government programs date back to the 1930s and even earlier. Scholars have unearthed evaluation efforts as early as 2200 B.C. Skipping some four millennia, econometric advances during the first half of this century, particularly in the 1920s and 1930s, laid the foundations for quantitative evaluations.[1] A massive study entitled *American Soldier* in 1945 and penal-rehabilitation studies in the 1950s are examples of early evaluations. The first congressionally mandated evaluation of a social program, a juvenile delinquency project, was recorded in 1962.[2] Nevertheless, the evaluation industry is largely a product of the past three decades. Before the 1960s, the tools of the trade were largely limited within the government to calculators, while the more sophisticated techniques found in the academic world were generally not applied to public policy issues. However, program administrators have always been concerned about the effectiveness of their efforts. Lacking adequate data bases, trained evaluation staffs, appropriate techniques, and technology to collect and analyze the relevant data, administrators frequently relied on the seat-of-their-pants guesstimates for the evaluation of the programs entrusted to them.

The Planning-Programming-Budgeting System

The impetus for evaluating social programs came in the mid-1960s from the Defense Department's Planning, Programming, and Budgeting System (PPBS). While this exercise focused on planning and estimating the most cost effective ways for achieving "the greatest bang for the buck," it also included evaluation components. The Defense Department developed PPBS as a tool to help managers achieve a more efficient administration. President Johnson was impressed with the technique and decreed that all other federal agencies develop and adopt their own PPBS.

---

[1] James J. Heckman, "Haavelmo and the Birth of Modern Econometrics", *Journal of Economic Literature* (forthcoming)

[2] William R. Shadish, Jr., Thomas D. Cook, and Laura C. Leviton, *Founders of Program Evaluation* (Newbury Park, CA: Sage Publications, 1985), pp. 20-27; Peter H. Rossi and Howard E. Freeman, *Evaluation: A Systematic Approach* (Newbury Park, CA: Sage Publications, 1979), pp. 21-29.

The objective of adopting PPBS was to improve government efficiency and allocation decisions by incorporating more accurate information into the decision-making process. Each agency needed to:

(1) define its organizational objectives;

(2) specify how money was being spent and the results achieved;

(3) identify policy options, together with estimates of the costs and impact of each; and

(4) design a system to provide all relevant information to decision makers in a timely manner.[3]

Although the Defense Department had developed considerable expertise in methods of analysis and planning that were integral to PPBS, most of the civilian agencies had, at best, limited staff devoted to such responsibilities. This led to technical difficulties in trying to implement PPBS in many civilian agencies. Some agencies saw the system as an imposition of no value to them, resulting in further resistance.[4]

The causes of PPBS' demise were numerous. For one, the introduction of PPBS without adequate preparation was bound to disrupt established practices and occasionally engender outright resistance. Agencies did not supply information fast enough to meet pressing needs.[5] The "ultimate" management planning system had put the cart before the horse. Linking planning with budgeting was fine in theory, but required evaluation results. Since sustained evaluative research began about the same time as PPBS, it arrived too late to do any good. This is somewhat ironic, since PPBS played an important role in shaping thinking about government programs that helped promote program evaluation. Due to the lack of theoretical underpinning, necessary data bases, as well as resources, the high hopes placed on PPBS never

---

[3] Alice Rivlin, *Systematic Thinking for Social Action* (Washington: The Brookings Institution, 1971), p. 3.

[4] Robert D. Lee, Jr. and Ronald W. Johnson, *Public Budgeting Systems* (Baltimore: University Park Press, 1973); Aaron Wildavsky, "Rescuing Policy Analysis from PPBS," *Public Administration Review*, March/April 1969, p. 193.

[5] Allen Schick, "A Death in the Bureaucracy: The Demise of Federal PPB," *Public Administration Review*, March/April, 1973, pp. 146-156.

materialized, and in less than a decade the system became history.

Expanding Evaluation Research

Although quickly abandoned, PPBS impressed upon policymakers and program administrators the need for evaluative research, improved managerial techniques, and emphasized the relative neglect of assessing social efforts. PPBS helped stimulate growth in applied social science research by generating a demand for such research and by establishing the administrative data that would make it possible.[6]

While the rise and fall of PPBS was both dramatic and quick, applied social science research grew steadily during the 1960s.[7] Aided in part by the PPBS experience, with its focus on the costs and benefits of programs, economists became the lead players in the burgeoning evaluation efforts. Incorporating the latest research methods and computer-aided statistical analyses, economics — along with social psychology, sociology, and political science —contributed to the establishment of the new discipline of evaluation. Literally thousands of evaluations were conducted. Although much of the research came from universities, which established their own courses and degree programs in evaluation, the demand was sufficiently great to support numerous private enterprises as well, both commercial and non-profit. Evaluators founded new journals and professional societies to disseminate their products, while established journals in the social and behavioral sciences devoted more of their pages to evaluation research. Consequently, evaluation became not only a new discipline but a new industry.[8]

Although economists were preeminent, the involvement

---

6 *Program Evaluation: Patterns and Directions*, Eleanor Chelimsky, ed, (Washington: American Society for Public Administration, 1985), p. 5.

7 Peter H. Rossi and Howard E. Freeman, *Evaluation: A Systematic Approach*, third edition (Beverly Hills, CA: Sage Publications, Inc., 1985), pp. 21-26.

8 Carol Hirschon Weiss, "Evaluating Social Programs: What Have We Learned," *Society*, November/December 1987, pp. 40-45; Albert D. Biderman and Laurie M. Sharp, *The Competitive Evaluation Research Industry* (Washington: Bureau of Social Science Research, Inc., 1972).

of other social and behavioral science researchers, working together with agency officials and politicians, not surprisingly led to the development of different evaluation methods. Policymakers sought comparisons between program costs and results. This can be accomplished with varying degrees of rigor by a variety of methods. However, evaluators have frequently devoted inadequate attention to the complexities resulting from multiple evaluation objectives. Differing objectives are best accomplished by alternative methodologies, each entailing different costs to gather the required information. In other words, evaluation came to require its own cost-benefit justification.

Three principal approaches emerged: microsimulation, experimentation (which will be treated here jointly with quasi-experimentation because of their close interaction), and qualitative studies. Each method has strengths and weaknesses and is likely to be the most appropriate research approach for specific types of situations. However, practitioners have too frequently considered their counterparts as rivals using inferior methodologies rather than as colleagues with alternative methods. This is particularly true of the rivalry between practitioners of social experimentation and quasi-experimental efforts.

Microsimulation

The development of national data bases and much greater access to affordable computing capability came at a very fortuitous time for social policy evaluators. As the Great Society programs of the Johnson administration were launched, there was understandably great interest in estimating the cost of contemplated initiatives and predicting their outcomes. Supporters wanted to find evidence that the programs were achieving their goals at an affordable price while detractors sought just the opposite. Administrators at all levels, as well as taxpayers, had an interest in the cost-effectiveness and impact of the programs. Microsimulation made the needed cost estimates attainable and became a significant tool of the policymaking process. Microsimulation could not and was not designed to either predict or measure program outcomes. By failing to provide some comparison of both costs and benefits, microsimulation could perform only part of an evaluation.

Prior to the development of microsimulation models, policy makers had little to go on in estimating the costs of potential program changes or proposed initiations beyond crude "back-of-an-envelope" calculations. Although the level of precision was ultimately grossly overstated, microsimulation provided at least the appearance of "hard numbers" for policy-makers' consideration.

Microsimulation employs a mathematical technique that allows the researcher to model the behavior of individuals, households, or firms. Generally, each unit is assigned a weight, based upon the number of like units it represents in the overall population from which the sample was drawn. Aggregating the weighted microsimulation results yields macro-level projections but not program results.

For example, to create a microsimulation projection involving taxation, researchers would use survey data on the characteristics of individual households, such as location, income, and composition (eg., the number of exemptions), together with knowledge of the tax codes. Armed with this information, an analyst can simulate tax liability resulting from the potential changes in tax codes. This, of course, is a highly simplified example. To illustrate further, using an actual case, the following steps would be required to analyze the effects of policy changes on food stamp program costs and caseloads, using Current Population Survey (CPS) data:

1. The data need to be modified from their original form to be compatible with the simulation model. Additional data on taxes and public assistance programs other than food stamps which affect food stamp eligibility (including supplemental security income, aid to families with dependent children, and state general assistance programs) would also be added to the CPS data.

2. The data files would have to be "aged," to reflect changes in household composition, incomes, the cost of living, and other variables from the base data period, when the data were gathered, to the time period for which the analysis was desired (obviously, the further out projections are attempted, the more critical the aging process, and the greater the potential for compounding errors into the models).

3. Taxes need to be simulated and added to the data

file, since after-tax income determines eligibility for food stamps.

4. Food stamp usage is simulated at least twice—once under existing rules, and once each under any proposed changes to the current rules—so that the results could be compared to measure the expected impact of the policy change.[9]

Most microsimulations employ static models, in which the behavior of individual actors is assumed to be constant, disregarding demographic, economic, or policy changes. More complex dynamic models attempt to estimate behavioral changes in response to policy changes, such as a labor supply response to a change in marginal income tax rates. Dynamic models are conceptually superior, but the inherent complexity in modeling even static behavior has prevented significant progress. Static models have therefore remained predominant. To further advance microsimulation as a tool for policy formulation, a panel of experts appointed by the National Research Council to evaluate microsimulation models recommended that federal agencies take "small steps" to develop dynamic microsimulation models.[10]

The data set most frequently used in microsimulations of tax and income transfer programs is the Current Population Survey (CPS), conducted monthly by the U.S. Bureau of the Census. Data sets are frequently supplemented

---

[9] Comptroller General of the U.S., *An Evaluation of the Use of the Transfer Income Model—TRIM—To Analyze Welfare Programs*, PAD-78-14 (Washington: U.S. General Accounting Office, November 25, 1977), pp. 10-12.

[10] National Research Council, Constance F. Citro and Eric A. Hanushek, eds, *Improving Information for Social Policy Discussions: The Use of Microsimu-lation Modeling* (Washington: The National Academy Press, 1991), p. 192.

[11] Sar Levitan and Frank Gallo, "Workforce Statistics: Do We Know What We Think We Know and What Should We Know?" (Washington: U.S. Congress, December 1989), pp. 2-5; Michael E. Borus, *Measuring the Impact of Employment-Related Social Programs* (Kalamazoo: The W.E. Upjohn Institute for Employment Research, 1979), pp. 57-64.

with additional data from other sources. Supplementation can be accomplished by matching data from different surveys that provide complementary information. Social Security or Internal Revenue Service files are potentially rich sources, although privacy considerations and statutory provisions greatly limit access to these files, and there are problems with the way data are reported. Social Security coverage excludes about 10 percent of all employment and IRS files provide little information on income sources and combine income on joint returns.[11]

Another means of supplementation, probably the least costly, is statistical imputation. This technique involves drawing data from a different survey sharing common variables with the main data base. Estimations of data derived by imputation vary from the quite simple to complex econometric models. For example, an imputation of capital gains income to the CPS file used IRS income (SOI) files. CPS families were split into tax filing units similar to those in the SOI and assigned to cells based on level of income and number of dependents. The analysis estimated the probability of a gain within each cell and the ratio of the gain to income for each randomly selected CPS unit.[12]

Imputation has advantages in that it is not limited to the donor data bases and is also generally cheaper and easier to accomplish than statistical matching. A potential disadvantage is that the variables added via imputation techniques are only as good as their estimation procedures, and, as with statistical matching, one can always question the validity of adding data, actual or estimated, from one unit of observation to another.[13]

---

12 Gordon H. Lewis and Richard C. Michel, eds, *Microsimulation Techniques for Tax and Transfer Analysis* (Washington: The Urban Institute Press, 1990), pp. 176-77.

13 Herman P. Miller and Roger A. Herriott, *Microsimulation: A Technique for Measuring the Impact of Federal Income Assistance Programs* (Durham, NC: Institute of Policy Science and Public Affairs, Duke University, 1977).

Simulation was first applied to welfare policy in 1969, when the President's Commission on Income Maintenance Programs attempted to estimate the costs of the Nixon administration's proposed Family Assistance Plan (FAP). Developed hurriedly to meet the demands of policymakers, the initial model, dubbed RIM (Reforms in Income Maintenance) lacked flexibility and documentation.[14] To improve on the initial effort, the Urban Institute in 1973 introduced TRIM (Transfer Income Model), the ancestor of many microsimulation models used today.[15] TRIM has enabled analysts to utilize the most recent survey data from multiple sources, to estimate the costs of a variety of tax and income transfer programs and proposals. In 1974, TRIM's designers moved on to Mathematica Policy Research, where they developed MATH (Micro Analysis of Transfer to Household). Funded by the U.S. Department of Agriculture's Food and Nutrition Service, MATH received wide acceptance and has been used by several executive departments and Congressional agencies.[16]

Microsimulation models played important roles in the 1975-76 food stamp debate, President Carter's 1977 aborted Better Jobs and Income proposal, the 1984-1986 tax reform debate, and most recently in welfare reform efforts. Despite its contributions, microsimulation has been severely criticized because projections were regarded as reality in some circles. Especially early on, the general lack of understanding among policymakers, not to mention policy analysts within the executive agencies and congressional staffs, left most users almost totally dependent upon estimates supplied by outside consultants. The estimates came out from the

14 Gordon H. Lewis and Richard C. Michel, eds, *op cit*, pp. 36-7.

15 Kenneth L. Kraemer, Seigfried Dickhoven, Susan Fallows Tierney, and John Leslie King, *Datawars: The Politics of Modeling in Federal Policymaking* (New York: Columbia University Press, 1987), pp. 33-62; Gordon H. Lewis and Richard C. Michel, eds, *op cit*, pp. 58-60.

16 National Research Council, *op cit*, pp. 110-111; Robert H. Haveman, *Poverty Policy and Poverty Research* (Madison, WI: The University of Wisconsin Press, 1987), pp. 110-43.

bowels of computers, but few policymakers understood how the estimates were derived or had any inkling about the assumptions made by the analysts to obtain the results.

Lack of validation was a major flaw of the microsimulation models. Given problems associated with sampling variability and input data, the analysts have failed to supply information about the reliability of the estimates. The National Research Council panel of experts identified the reasons for this shortcoming. First, the analysts could not distinguish between errors in their projections due to the behavior of program participants and the overall changes in the economy. Second, the analysts lacked incentives to develop measures of uncertainty because the policymakers were interested in a single number. The reliability of the estimates remained questionable.[17]

Under the circumstances, it was inevitable that the messengers were blamed for the message they delivered. Use of microsimulation in a 1975-76 food stamp debate raised questions about the objectivity of the technique. A rapid rise in the food stamp program's caseload prompted the Ford administration to propose significant cuts in the program, requiring program participants to "purchase" their food stamps at some ratio of their face value, rather than to continue receiving them at no cost. The Department of Agriculture's Food and Nutrition Service, which administered the food stamp program, contracted with Mathematica Policy Research to estimate both the budgetary impact of the proposed changes and their impact on program participants.[18]

Skeptics in Congress, who questioned the findings and the objectivity of their operators, sought a second opinion. They turned to the independent U.S. General Accounting Office to test possible bias in the various food stamp models. The GAO report, released in 1977, faulted the modelers for using questionable techniques. The GAO analysts acknowledged that policymakers frequently demand exact estimates, but this did not prevent them from criticizing the modelers for the frequent reporting of point estimates rather than a range of estimates. Analysts supply a single number ostensibly to

17 National Research Council, *op cit*, pp. 3-5, 231-264.
18 Kenneth L. Kraemer, *et al.*, *op cit*, pp. 110-111.

avoid "confusing" policymakers, but this practice lends a misleading aura of "scientific" accuracy to their findings, a failing often shared with experimentalists' and quasi-experimentalists' methods (discussed later). The GAO report noted, according to one estimate, that adjusting the data to account for income underreporting could result in a 10 percent reduction of eligible households. The report also stressed that the validity of the original survey data are always subject to dispute, a potential problem that is compounded by the frequent massaging of the data in simulation models. Finally, the GAO report cautioned that microsimulations are inappropriate for long-range projections (eg., estimates beyond four or five years), and, if used, it is incumbent that analysts clearly indicate the weaknesses of the projection.[19]

Certainly, ideological concerns about who was running the microsimulations were a factor in requesting the GAO report. However, even assuming that the modelers are honorable people of the highest integrity, they must still make a large number of judgments along the way, nearly any or all of which could bring bias into the estimates. Data limitations also raise serious concerns. For example, in analyzing a transfer program, deriving a sample of AFDC recipients from the CPS is likely to yield a small sample size, and the underreporting of income is also a problem. Also, data from other sources are frequently "matched" with the files from the principal source, generally providing the opportunity for error with each addition. Finally, the behavioral constraints built into the model are considered by many to be the most difficult hurdle faced by microsimulation modelers. Even when empirical estimates of behavior exist, such as the labor supply elasticities, the resulting estimates will often be sensitive to the selections the modeler chooses. The impact of changes in wage levels upon the number of persons willing to work offers one illustration.[20]

---

19 Comptroller General of the U.S., *An Evaluation of the Use of the Transfer Income Model—TRIM—to Analyze Welfare Programs*, Report to Congress, PAD-78-14 (Washington: U.S. General Accounting Office, November 25, 1977), pp. 65, 90-93.
20 Comptroller General of the U.S., *Ibid.*, pp. 61-89; Robert Haveman, *op cit*, pp. 226-229.

Discounting the inherent problems of microsimulation models, the method rapidly gained credibility as an appropriate tool for estimating the costs of changes in social programs. Cutbacks in federal evaluation and research funding in the 1980s reduced the use, development, and refinement of microsimulation models. Microsimulation models nonetheless remain one of the evaluation industry's first-line tools for estimating costs of contemplated initiatives, and ill-advised or not, Congress and other policymakers routinely demand and use their estimates in considering legislation. Although the microsimulation models have been developed at public expense, the modelers seem to have kept tight control of their products. "As a result, only a handful of people know how to - or can afford - to use them."[21]

Experimentation and Quasi-Experimentation

Microsimulation can provide estimates of the cost but not the other effects of a policy, especially the impact of a program on participants. In contrast, social experimentation focuses on the impact of programs or demonstration projects already in place. While microsimulation attempts to provide estimates on "what it would cost," experimentalists focus on providing information on "what did happen." If one accepts the experimentalists' estimates, the critical issue is to determine whether the measured outcome is worth the cost.

Social experiments attempt to duplicate the control methods used in, say, agricultural experiments, giving researchers faith (though they might prefer the word "confidence") that the impacts they observe are indeed the result of a designated treatment. The counterfactual result—what would have happened if a person had not participated in a program—can, in theory, only be ascertained if a control group identical to the treatment group is part of the experiment. Random assignment to control and treatment groups allows the researcher to assume that the two groups were identical prior to treatment.

21 Julie Kosterlitz, "Education Guesswork," *National Journal*, October 10, 1991, p. 2412.

During the last decade, many policy analysts increasingly have maintained that a reliable estimate of the impact of social programs can only be obtained from controlled field experiments, in which participants and nonparticipants are randomly selected to test the impact of programs under consideration. This claim has led several prominent analysts to conclude that based on recent methodological developments, "a consensus seems to be emerging that... random assignment should be the *sine qua non* of future evaluations."[22] Similarly, "the argument for systematic experimentation is straightforward: Information necessary to improve the effectiveness of social services is impossible to obtain any other way."[23] If the claims were realized, they would change the way social policy research is conducted and utilized. This view is not universally shared. Another social scientist warned, however, that "We are in danger of pretentious scientism, which defensively claims more precision and control than has been achieved".[24] Although uttered two decades ago, the admonition holds equally today. More recently a prominent economist has argued "...advocates of randomization have overstated their case....Bias introduced by randomization is a serious possibility."[25]

Nonexperimental evaluations differ from experimental evaluations by not using a randomly assigned control group. In the absence of random assignment, quasi-experiments typically construct a control group from a population judged to be comparable to those who received the treatment. Quasi-experimentalists employ statistical techniques to disentangle program effects from other sources of differences in outcomes. Quasi-experimental studies have an advantage over experiments in that they are less intrusive on the operation of a program and are also less costly. As a result, pro-

[22] Isabel V. Sawhill, "Poverty in the U.S.: Why Is It So Persistent?" *Journal of Economic Literature*, September 1988, p. 1094.

[23] Alice Rivlin, *Systematic Thinking for Social Action* (Washington: The Brookings Institution, 1971), p. 108.

[24] Donald T. Campbell, "Considering the Case Against the Experimental Evaluations," *Administrative Science Quarterly*, No. 1, 1970, p. 1094.

[25] James J. Heckman, "Randomization and Social Policy Evaluation" (Cambridge, MA: National Bureau of Economic Research, Technical Working Paper No. 107, July 1991), p. 3.

gram administrators are more likely to participate in a quasi-experimental evaluation. Also, sample size is less apt to be a problem, although it may present a technical challenge when matches are sought. Proponents argue that results from a rigorously conducted quasi-experiment are likely to be as robust as the results obtained from experimental evaluations. Experimentalists counter that nonexperimental methods produce inherently unreliable estimates and that only controlled social experimentation yields meaningful results.[26] The allocation of federal and foundation social research finding has reflected their view. Although estimates vary, depending upon who is counting and what is counted, a compendium of 63 social experiments conducted between 1967 and 1990 carried a price tag of $916 million (1991 dollars). This total excluded 13 experiments for which cost data were not available and 28 others that were in progress as of August 1990.[27] Another estimate placed the costs of experiments initiated between 1968 and 1975 at $1.5 billion (1991 dollars).[28] Estimates of federal expenditures for nonexperimental (including qualitative) evaluations are not available.

The Case for Randomization

The rationale for social experimentation is straightforward. Reliable judgments about the impact of social programs are difficult to obtain because direct causal effects on individual participants are rarely, if ever, observable. If it were possible to directly observe not only the post-program characteristics of participants, but also the course their lives would have taken if they had not participated, then the program's effects could be directly measured. Researchers attempt to account for unobserved factors, but no methodology can truly

---

26 Robert J. LaLonde, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *Evaluation Studies Annual Review*, Vol. 12, William R. Shadish, Jr. and Charles S. Reichardt, eds (Newbury Park, CA: Sage Publications, 1987), p. 604; Thomas Fraker and Rebecca Maynard, "The Use of Comparison Group Designs in Evaluations of Employment-Related Programs" (Princeton, Mathematica Policy Research, 1985).

27 David Greenberg and Mark Shroder, *Digest of the Social Experiments* (Madison, WI: Institute for Research on Poverty, Special Report #52, May 1991), p. III.

28 Robert Haveman, *op cit*, pp. 183-185.

measure what did not happen. Judgments about how any specific program participants' lives might have been different if they had not participated in a program must always remain speculative.

The advantage of randomization is that it subjects uncontrollable sources of uncertainty to a probability calculus. Random assignment is sometimes said to ensure that differences between the program enrollees and nonenrollees are caused solely by program effects. That is neither its effect nor its intent. R.A. Fisher, whose work contributed to the acceptance of randomization and experimental design, observed that extraneous causal factors are "always strictly innumerable." As a result, he continued: "...whatever degree of care and experimental skill is expended in equalizing the conditions, other than the one under test, must ... always be to a greater or lesser extent incomplete, and in many ways grossly defective."[29]

As in nonexperimental models, any two random assignments will produce different enrollee and nonenrollee groups, and can be expected to produce different estimates of program impact. But while repetition of a random assignment experiment produces a range of estimates, in theory the average of this range can be expected to settle at the true value of the program impact, assuming a sufficient number of sample studies have been collected. That is, randomization assures that differences between enrollees and nonenrollees are caused by either program effects or sampling error. Randomization-based inference provides unbiased estimates of mean treatment effects, when median effects may be more significant for policy formulation and evaluation of ongoing programs.[30]

Selection Modeling

Lacking direct observations of program impact on individual participants, policy analysts estimate the impact of the enrollment by comparing the post-program characteristics of selected groups of participants and nonparticipants. In a

[29] R.A. Fisher, *The Design of Experiments* (Edinburgh and London: Oliver and Boyd, 1935), p. 19.

[30] James J. Heckman, "Randomization and Social Policy Evaluation," Technical Working Paper No. 107 (Cambridge, MA: National Bureau of Economic Research, July 1991), p. 3.

laboratory, researchers exercise a very substantial degree of control, but this power is much more circumscribed in social policy experiments. Experimentalists intervene in the selection of enrollees in order to separate a program's effects from other sources of differences between participants and nonparticipants. If the groups were identical except for enrollment or nonenrollment in the program being evaluated and if no alternatives for participation in other programs were available to the controls, then differences in outcomes between the two groups could be unambiguously attributed to the effects of the program. But the effects of program participation are typically only one among many sources of differences between participants and nonparticipants. The latter, for example, might enroll in a different program and receive similar treatment than the participants. Consequently, the differences between the two groups that are attributed to the effects of the program under test may in fact also reflect, to an unknown degree, all of the other vicissitudes of life. This problem, which is referred to as "selection bias," has been a dominant methodological concern of public policy analysts since the 1960s.

To the extent that enrollees' propensities to utilize different services can be predicted from their observable characteristics, such as age or income, it is possible to statistically control for this self-selection by creating matching groups of enrollees. However, the differences between enrollees that affect their likelihood of needing particular services frequently are not observable in this way. To this extent, nonexperimental matching of enrollees on the basis of their observed characteristics is insufficient to create comparable groups for evaluation purposes.

In the absence of random assignment, quasi-experiment modelers typically construct a control group from a population judged to be comparable to those who received the treatment. They match, as far as possible, program enrollees and nonenrollees with respect to measurable extraneous causal factors, but do not employ random assignment. The principal difficulty with quasi-experimental studies is that they fail to quantify the impact of unknown extraneous causal factors. For example, analysis of the impact of training programs provided under the Comprehensive Employment and Training Administration, a program that provided jobs or training to low income persons, matched participant files

gathered in the Continuous Longitudinal Manpower Survey (CLMS) with a comparison group selected from the March Current Population Survey. Both groups were first matched with Social Security Administration earning files. Pre-program earnings were then used in the matching process, along with demographic, socio-economic, and work history variables in the CPS and CLMS surveys, while post-program earnings formed the basis for measuring the impact of training. Different matching schemes were tried, using slightly different variables and altering the relative weights given to each. The results differed depending upon the matching method chosen, so that it was unclear whether the study's results reflected program effect or the matching methodology. Consequently, all methods were judged to be at least somewhat inadequate for the desired task.[31] To overcome the difficulties associated with matching CLMS and CPS data, the Urban Institute researchers constructed their own quasi-experimental control group using eligible nonparticipants in evaluating an AFDC training program. The analysts matched the treatment with the control groups using age, sex, race, and family size.[32]

The Urban Institute designed another promising quasi-experimental model for evaluating the Washington State Family Independence Program, a modified version of the 1988 welfare reform JOBS program. Rejecting, in this case, a random assignment approach as being "infeasible or unwise," the Urban Institute matched comparison sites. In selecting the matched communities the analysts chose relevant characteristics, including earnings, unemployment, out-of-wedlock births, and placements of AFDC recipients. In

---

31 Office of Program Evaluation, U.S. Department of Labor, "The Impact of CETA on Participant Earnings: Entrants During the First Half of 1975, Continuous Longitudinal Manpower Survey, Working Paper #1" (Rockville, MD: Westat Research, Inc.), Contract No. 23-24-75-07, January 1980.

32 Demetra Smith Nightingale et al., *Evaluation of Massachusetts Employment and Training (ET) Program* (Washington: The Urban Institute, 1991); Lee Bawden, "Comparison Group Design" paper presented at American Enterprise Institute Conference on Child Welfare, February 20, 1991, p. 6.

addition to obtaining outcome results the analysts obtained pre-program differences to be compared with later follow up. The intent was to minimize unexplained variables and thus provide a valid basis for estimating the net impact of the Washington state program.[33]

As they were launching the first major social experiments, policy analysts turned to the problem of developing reliable statistical methods for drawing causal inferences from nonexperimental data when selection bias is present. The reliability of nonexperimental evaluation models substantially depends on the comparability of participants and nonparticipants. As noted, the reliability of nonexperimental program evaluations was called into question by the divergent results of nonexperimental program evaluations authorized under Comprehensive Employment and Training Act (CETA) of 1973.[34] The evaluations were all based on data from the same sources, the Continuous Longitudinal Manpower Survey, (a longitudinal sample survey of CETA enrollees designed specifically for the purpose of evaluating CETA) and the CPS (the source of the matched control group). The methods that analysts had utilized for constructing the control group were widely interpreted to provide no basis for judging which, if any, of the various findings was more reliable.

Simultaneously, researchers at Princeton University and Mathematica Policy Research reported comparisons of experimentally- and nonexperimentally-based impact estimates for the National Supported Work Demonstration, a random assignment experiment conducted between 1975 and 1979 to test the effects of structured work experience on long-term

---

33 Lee Bawden and Freya L. Sonenstein, "Quasi-Experimental Designs: Sometimes Preferable, Sometimes Necessary (Washington: The Urban Institute, 1991) pp. 7-9; Demetra Smith Nightingale et al., *Services to Clients in Washington State* (Washington: The Urban Institute, July 1991), pp. 1-16.

34 Howard S. Bloom and Maureen A. McLaughlin, *CETA Training Programs: Do They Work for Adults?* (Washington: Congressional Budget Office, July 1982); Laurie J. Bassi, "The Effect of CETA on the Postprogram Earnings of Participants," *Journal of Human Resources*, Fall 1983, pp. 539-556; Westat, *Summary of Net Impact Result* (Rockville, MD: Westat, 1984).

welfare recipients, ex-addicts, ex-offenders, and young high school dropouts. These comparisons confirmed to the satisfaction of the researchers that the divergence of the results of nonexperimental evaluations from those of experimental evaluations could be of major practical significance.[35]

The critical view about the unreliability of quasi-experimental evaluation apparently has been effectively challenged. The critics reject the claim that "experimental methods are the only way to accurately evaluate the impact of manpower training programs on outcomes...."[36] They claim that they could derive the same results with quasi-experimental methods by using alternative estimation procedures. The fact that the CETA evaluations obtained widely different results was due to different assumptions that could have been tested. The Job Training Partnership Act evaluation (discussed later) is using experimental and nonexperimental methods. This exercise may offer some insights into the controversy but will, of course, not resolve the debate.

From Sunday's Sermon to Monday's Work

As the proponents of social experimentation have turned to the realities of implementing their designs, numerous quandaries have arisen. Experiments frequently cannot be implemented as intended, and their results are always more highly subject to interpretation than some analysts have suggested.

---

35 Robert J. LaLonde, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," Working Paper no. 183, Industrial Relations Section, Princeton University, October 1984; Thomas Fraker and Rebecca Maynard, "The Use of Comparison Group Designs in Evaluations of Employment-Related Programs" (Princeton, NJ: Mathematica Policy Research, 1985).

36 James J. Heckman and V. Joseph Hotz, "Are Classical Experiments Necessary for Evaluating the Impact of Manpower Training Assessment? A Critical Assessment," Industrial Relations Research Association. *Proceedings of the Fortieth Annual Meeting*, Barbara D. Dennis, ed (Madison, WI: The Association, 1987), p. 299; — "Choosing Among Alternative Nonexperimental Methods of Estimating the Impact of Social Programs," *Journal of the American Statistical Association*, December 1989, pp. 862-863.

As the debate over welfare reform picked up steam in the latter half of the 1960s, the concept of a negative income tax (NIT) received serious attention. The theoretical advantages of such a system were many. For one, proponents hoped that much of the bureaucratic structure responsible for providing the varied forms of income support and in-kind benefits could be replaced with a single cash grant administered by a single agency. Another advantage of NIT would have been the reduction in disparities in benefit levels across states. Finally, NIT proponents argued that the work disincentives of income support programs could be reduced by having lower implicit taxes on recipients' earnings.[37] It was this last advantage that framed the key research question concerning the NIT in the early days of its discussion.

The New Jersey Income Support Experiment. To gain insights into the impacts of a guaranteed income program upon work behavior, the Great Society's antipoverty agency, the Office of Economic Opportunity, launched the first major income maintenance experiment, the New Jersey Graduated Work Incentive experiment, in 1967. The project illustrates many pitfalls that exist in experimental design and execution, and should have alerted proponents to avoid excessive claims for their approach.[38] The analysts assembled treatment and control groups in four cities to test the reactions of participants to different combinations of guaranteed income levels and marginal tax rates. The experiment cost $26 million (1991 dollars) and ran four years; more than two additional years were taken up in planning and post-program analysis. When the final report was published in July 1974, the authors concluded that there was no significant pattern of decreased work effort associated with the guaranteed annual income program.[39]

---

[37] Christopher Green, *Negative Income Taxes and the Poverty Problem* (Washington: The Brookings Institution, 1967); Milton Friedman, *Capitalism and Freedom* (Chicago: The University of Chicago Press), 1962, pp. 190-95; James Tobin, Joseph A. Pechman, and Peter M. Meiszkowski, "Is a Negative Income Tax Practical?" *Yale Law Journal*, November 1967, pp. 1- 27.

[38] Sar A. Levitan and Gregory Wurzberg, *Evaluating Federal Social Programs: An Uncertain Art* (Kalamazoo, MI: The Upjohn Institute for Employment Research, 1979), pp. 26-30.

[39] David Kershaw and Jerilyn Fair, *The New Jersey Income-Maintenance Experiment* (New York: Academic Press, 1976).

The overhead of the New Jersey experiment is a useful starting point for an examination of systematic experimentation costs. Two-thirds of the total cost of the project were allocated for assembling the research staff, selecting the sample populations for the control and experimental groups, conducting interviews, and preparing the final report. The experiment proved more expensive to set up than anticipated because of the difficulty in securing an adequate racial balance in each experimental income-range sample. Thorough and controlled documentation required frequent interviews and reporting.

Nonetheless, obvious flaws could not be avoided. In fact, inherent in the very design, so carefully constructed, was an observation process that interfered with normal behavior. The experimenters did not analyze the ramifications of extensive observations, only noting that the experimental group was more adept at filling out the income forms than the control group because the former completed them more frequently — monthly compared with quarterly by the control group.

Omission of an influential variable is a potential flaw that can completely vitiate the experimental approach. The primary objective of the New Jersey experiment was to test how different minimum income levels and marginal tax rates affected the incentive of participants to work. The evidence showed no significant pattern of persons dropping out of the labor force in response to high marginal tax rates support levels. The experimenters concluded that incremental changes in support levels and marginal tax rates would not cause people to deliberately cut back their work effort.[40] However, changes associated with the level of in-kind benefits probably would.[41]

Perhaps the most serious drawback to systematic social experimentation concerns the assumption that experimental conditions are attainable in a social setting. People

---

40 David Kershaw and Felicity Skidmore, "The New Jersey Graduated Work Incentive Program" (Washington: Office of Economic Opportunity, July 1974).

41 Robert I. Lerman and Alair A. Townsend, "Conflicting Objectives in Income Maintenance Programs" (*American Economic Association Proceedings of Annual Meeting*, May 1974), pp. 208-210.

and the societal setting cannot be as easily manipulated as peas in a laboratory. Three uncontrollable extraneous factors influenced the New Jersey experiment. The first outside influence was anticipated: participant and community attitude towards the experiment. There was reluctance to cooperate with "do-gooders" studying the poor; militant community opposition surfaced when the experiment began. Program designers were also concerned about participant reactions to varying benefits, as well as community and control group acceptance of an experiment that excluded benefits altogether for the latter group.

A second serious disturbance in the experiment was the January 1969 change in the New Jersey state law that qualified families headed by unemployed fathers to receive Aid to Families with Dependent Children. This change, also beyond the control of the evaluators, altered an important experimental precondition. One of the reasons New Jersey was selected as a site for testing the negative income tax was the lack of any welfare program for unemployed fathers (or plans for one). But the world did not stand still. In order to minimize the competition and overlap between the demonstration and the newly installed benefits, the experimenters added another group of participants who received higher guaranteed support payments. This raised the costs of the experiment and necessitated payments in excess of the poverty guidelines.

A third problem encountered by the New Jersey experimenters involved changes in the political context which destroyed the analogy of a social laboratory. The New Jersey experiment encountered just this kind of problem, as political reality overtook the experiment. In August 1969, when the project had been underway barely a year, the Nixon Administration unveiled its Family Assistance Plan, a welfare reform proposal which featured a negative income tax provision. Since the New Jersey experiment was then the most prominent source of empirical data about the relationship between a negative income tax and work behavior, the House Ways and Means Committee of the United States Congress called the project administrators to testify on their findings. At their first appearance, they responded to the congressional inquiries in general terms. Thereafter, they released their preliminary findings that supported the family assistance plan in principle. The report opened the experiment to close scrutiny

by the General Accounting Office and the Senate Finance Committee. The latter attempted to obtain confidential records of participating families — a potentially disastrous action from the experimenters' point of view. The disclosure issue died, but the project administrators were hard put to comment on specific questions without taking sides in the debate.

Project officials noted that as experiments became more relevant to current political decisions, legislative inquiries were more likely to pose sensitive issues and threaten the success of the experiment. Observers of the New Jersey experiment commented that serious complications may arise not merely from bad judgment but even more from plain bad luck.[42] Critics of experimentation charged "... the classical model [is] costly, lengthy in its execution, and uninterpretable, in any prescriptive sense, for policy."[43]

Selecting a Probabilistic Sample. A common initial problem that evaluators of ongoing programs face is the need to secure the cooperation of program administrators. Experience suggests the difficulties that may disrupt the implementation of an experiment. Given the opportunity, program administrators are more likely to refuse to participate in an experimental study than in a comparable nonexperimental study. The reasons for this differ from case to case, but may include a desire to avoid the disruption of program operations or recruitment that the experiment would entail, ethical objections to random assignment in the provision of program services, and aversion to scrutiny by an outside group. The potential severity of these problems is illustrated by the experience of the U.S. Labor Department's Job Training Partnership Act evaluation, which to date has cost an estimated of $28.3 million (1991 dollars). The program is largely devoted to training unskilled, low income, unemployed persons. The original research design for this study called for a random assignment in a probability sample of 20 of JTPA's 630 service delivery areas (SDAs). This plan had to be aban-

[42] Peter H. Rossi and Katherine C. Lyall, "An Overview Evaluation of the NIT Experiment," in *Evaluation Studies Review Annual*, Thomas D. Cook et al, eds (Beverly Hills, CA: Sage Publications, 1978), Vol. 3, pp. 412-428.

[43] Eleanor Chelimsky, ed, *Program Evaluations: Patterns and Directions* (Washington: American Society for Public Administration, 1985), p. 13.

doned, however, because 74 percent of local program administrators refused to participate in the study. A total of 229 sites were contacted about their interest to participate in the study. Of these, 170 refused to participate and another 43 were determined to be not appropriate.[44] Ultimately, the experiment's planners were able to obtain, at a cash payment of four times the original intended cost, a nonprobabilistic sample of 16 SDAs which they claimed is generally "representative of the diversity" of SDAs in the JTPA system. Considering the wide variability in program operations at different SDAs, the validity of the claim is questionable.

The analysts' inability to implement the initial probability sample design vitiated the experiment's ability to produce representative estimates of JTPA's training impact.[45] Whether the programs operated in these 16 sites are like those in the 170 who declined to participate is doubtful, because administrators of subpar programs are more likely to avoid scrutiny. The experiment has therefore only speculative applicability to the entire JTPA system. Moreover, the samples obtained from each SDA were too small to allow valid site-by-site analysis of the program's impact on the treatment groups.[46] If the problem of local administrators' refusal to participate in experimental evaluation is to be avoided or mitigated in the future, the evaluators need to gain an understanding of the institutional context in which the programs operate or they need to acquire the power to exercise sanctions to secure cooperation.

Even when administrators participate in experiments, they may deny or impede access to essential information, change their practices in ways that reduce the reliability of

---

44 Fred Doolittle and Linda Traeger, *Implementing the National JTPA Study* (New York: Manpower Demonstration Research Corporation, April 1990), pp. 32-39, 92.

45 Burt S. Barnow, "The User and Limits of Social Experimentation," Industrial Relations Research Association, *Proceedings of the 40th Annual Meeting*, Barbara D. Dennis, ed (Madison, WI: The Association), p. 280.

46 V. Joseph Hotz, "Designing Evaluations of the Job Training Partnership Act," in *Evaluating Welfare and Training Programs*, Charles F. Manski and Irwin Garfinkel, eds (Boston, MA: Harvard University Press, 1992), pp. 110-112.

data obtained from the experiment, or act in other ways that increase the cost or reduce the analytical value of experimental data. No matter how careful the design of an evaluation may be, its implementation involves many potential pitfalls. For example, relying on "hired hands" to do the interviewing may lead to data corruption if the interviewers perceived that they are relegated to do the "dirty work." Of course, the entity performing the experiment may decide to use its regular staff to perform the chore. However, such an approach may involve additional costs.[47]

The Black Box

The data of social experiments, at least initially, are supposed to measure the simple observed relationships between treatments and effects, without uncovering the underlying causes that produce the effects. This strategy sometimes is referred to as "black-box" science, defined by Webster's dictionary as an "electronic device whose internal mechanism is hidden or mysterious to the user." Methodological arguments for social experimentation most often construe it as a model of this approach. Experimental evaluations of complex programs such as the Job Corps, a residential skills training and educational program for disadvantaged youth, yield data on their average effects, but no direct information about the unobserved causal components through which the programs produces these effects. The averages do not provide any insights about whether the measured outcomes were due to the removal of enrollees from their presumably debilitating environments, the training and education offered, or other program components. However, it is critically important to know the impact of each program component in order to fashion improved programs, and a "black box" approach fails this test.

Elusiveness of the Target Population. The black-box approach produces estimates of the average effect of the program under test on participants in the experiment, but frequently the population outside the program is also of great interest to policymakers. When contemplating the expansion of an existing program, they may desire information about its impact on

[47] Eva Lantos Rezmovic, Thomas J. Cook, and L. Douglas Dobson, "Beyond Random Assignment," *Evaluation Review*, February 1981, pp. 56-59.

marginal new enrollees or on all eligible persons. To the extent that these groups differ from the current enrollees, experimental results will be inapplicable for the purposes of estimating the effects of an expanded program. Conversely, even estimates of net program impact on current enrollees may be difficult to obtain if the number of this group at any given site is small. In such cases, analysts frequently find it necessary, in order to create an adequate control group, to enroll additional participants in the experiment and thereby raise the cost of the experiment. To the extent that these new subjects differ from the original program enrollees, experimental results again will produce biased estimates of the program's impact on its original enrollees. In all of these cases, in addition to raising costs and complicating administrative procedures, estimates of program impact on marginal enrollees or on other special groups can be obtained only by invoking more assumptions than is proper under the black-box approach.

Even when it is possible to obtain estimates for the target population, some of the observations called for in the experimental research design typically cannot be obtained. This problem, referred to as "nonresponse bias," is a recurrent source of uncertainty about the applicability of experimental results to the target population. It often arises from a refusal of program sites to participate in the experiment and from the attrition of experimental subjects, or their refusal to volunteer information.

Experiments rarely produce measurements of the entire target population. Normally, they are confined to sample surveys of the target population. In these cases analysts observe the behavior of the sample and make inferences from the sample to the whole population. The degree to which these inferences are justified depends in large part not only on the size of the sample, but on the way the sample is selected and the applied statistical inference procedures. However, the designers of social experiments may choose to forgo a probability sample for reasons of convenience or cost. But even when social experiments are intended to draw a probability sample, the research design may be difficult to implement. In all such cases, the experimental data offer no secure formal quantitative basis for generalization from the sample to the target population. These problems are not unique to experimental studies, but they frequently are more severe in experi-

ments than in comparable nonexperimental studies because, as noted earlier, experimental studies involve a greater degree of intervention than quasi-experiments. As a result, even if they are internally valid, experiments would provide a weaker foundation for generalization to the target population, sometimes referred to as "external validity," than would nonexperimental methods that had employed probability sampling techniques.

This problem arises especially in experimental designs that include probability sampling of program sites when local program administrators refuse to participate in the evaluation. If this refusal were to occur randomly, its only consequence would be the loss of precision that is always associated with reductions in sample size. However, as the JTPA evaluation illustrates, the refusal by program administrators to participate normally occurs in a nonrandom manner.[48]

Attrition. All social evaluations are affected to some degree by problems of attrition. As is the case with the refusal of administrators to participate in experiments, purely random attrition would introduce no systematic bias into estimates of program effects. But attrition normally occurs nonrandomly. While administrators' refusal to participate affects the validity of inferences from participating sites to the overall national program, attrition of enrollees within participating sites affects the estimated reliability of the program's effects on the sample in the participating sites. Analysts frequently attempt to estimate the extent of the attrition problem, but in doing so they must employ the same techniques of matching according to observable characteristics that proponents of social experiments have criticized in nonexperimental and quasi-experimental studies.

Randomization may in other ways make it impossible to obtain a sample that is representative of the target population. Subjects who choose to enter experimental programs may incur some opportunity cost by doing so, and also accept a risk of being assigned to a control group that does not receive services. Potential subjects can be expected to weigh this risk differently, and the more highly risk-averse among

[48] Fred Doolittle and Linda Traeger, *op cit*, pp. 89-93.

them may be less likely to choose to enter the program than others. Because risk-aversion is a basic dimension of personality that has broad effects on behavior, randomization itself may make it difficult to make inferences from the effects of a program on a self-selected risk-acceptant group to the broader target population.

The Need to Generalize to Untested Treatments. Because the black-box approach eschews the analysis of underlying, unobservable causal processes, it cannot provide an adequate basis for generalization from the effects of the tested treatments to alternative programs, even very similar efforts. But policymakers rarely are interested only in the effects of one specific program component; they seek to choose among a range of alternatives. They require estimates of the effects of these alternatives, but "the very rigor of social experimentation limits the policy relevance of the results." The only alternative that black-box science offers, "one experiment per social program," generally is incapable of anticipating the whole range of options that policymakers may wish to contemplate. If analysts would try to estimate the total options of interest, the cost of conducting the experiment may prove to be prohibitive. A sociologist charged that experimentalists have made "no effort ... to penetrate the black-box of causation," placing emphasis "on elegant manipulation rather than interpretation of narrative and quantitative information." Another analyst charged that experimental evaluations do not provide "insights on noneconomic rewards, such as the psychic gains that people receive from earning their own income."[49]

The black-box model does not prescribe procedures involved in implementing social experiments. Most analyses of the major social experiments have been designed for the purpose of fitting the data to microeconomic models.[50] To the extent that they depart from the black-box model, analysts

[49] Alice H. Munnell, "Lessons from the Income Maintenance Experiments: An Overview," *New England Economic Review*, May/June 1987, p. 41 (views by Richard Elmore, Lee Rainwater, and Charles Murray, respectively).
[50] Orley Ashenfelter and Mark W. Plant, "Nonparametric Estimates of the Labor-Supply Effects of Negative Income Tax Programs," *Journal of Labor Economics*, January 1990, pp. 396-415.

of social experiments resort to the same tools that they often reject in nonexperimental studies.

Biases Arising From the Limited Duration of Experiments. The effects of a limited-duration demonstration project may not be generalizable to an otherwise identical permanent program.[51] Program participants' expectations reflect in part the duration of the treatments to which they are exposed, but theory and experience suggest that human behavior is strongly conditioned by expectations of the future. Many of the behavioral responses of interest in social experiments, including marital stability and labor market attachment, may be conditioned by enrollees expectations of the duration that program benefits would be available to them.

Feedback Effects. Among the extraneous causal factors that influence program outcomes may be variables strongly determined by the feedback effects of the social policy environment. In particular, the effects of a nationally or regionally established program may differ from the effects of an identical small-scale experimental program. For example, the effects of an experimental job training program can be expected to strongly depend on local labor market conditions, and yet if the experimental program were introduced on a permanent basis and were open to wider participation, the new program could itself affect labor market conditions in a way that could change its effects on the participants. Similarly, the impact of services designed to help welfare recipients to seek economic self-sufficiency may depend upon the level of the state's cash support. In such cases, experimental estimates of program impact, even if reliable for the small-scale trials on which they are based, would not be generalizable to otherwise identical large-scale programs. Moreover, considering the high costs of controlled experiments, they are likely to be limited to a few areas or limited samples which may not be representative of the total program or population being studied.

Hawthorne Effects. Finally, experimental results may be misleading because the behavior of participants is likely to be influenced and modified by the knowledge that they are being observed — the Hawthorne effect. The anticipation that their behavior may be used as evidence to justify changes in public policy, may lead subjects to behave differently in experimental situations than they would if they were

unobserved. Therefore, conclusions based on experimental results may not be warranted for policies carried out under normal conditions.[52] The methods that have hitherto been used to assess the severity of Hawthorne effects depend on assumptions similar to ones that have been criticized in non-experimental studies. However, some evaluators have disputed the significance of Hawthorne effects in experimental outcomes. They reasoned that "individuals respond rationally to the incentives they face whether experimental or not."[53]

The Elusiveness of Consensus

Notwithstanding the claims of proponents that social experiments offer a rational basis for consensus about the effects of programs, the interpretation of experimental results has itself often been a source of controversy. Findings of experiments conducted under full employment may not apply to slack labor markets. Similarly, reactions of one community may differ from another depending upon social mores.[54] Finally, results obtained from temporary demonstration projects with restricted scopes may have limited application to broader programs with no expiration date, making replication problematic.

Two recent controversies over the interpretation of experiments from the fields of income maintenance and criminal justice illustrate the pitfalls of a rush to judgment. The competence and qualifications of the participants in the controversies are not in question; in each case all are reputable scholars with extensive experience in the analysis of social experiments, but in neither case have the issues in dispute been resolved. The disagreements have involved matters of method as well as of fact. The following two examples are

---

51 Gary Burtless and David Greenberg, "Inferences Concerning Labor Supply Behavior Based on Limited-Duration Experiments," *American Economic Review*, June 1982, pp. 488-497.

52 Alice Rivlin, *Systematic Thinking for Social Action* (Washington: The Brookings Institution, 1971) p. 116.

53 Gary Burtless and Larry R. Orr, "Are Classical Experiments Needed for Manpower Policy," *The Journal of Human Resources*, Fall 1986, p. 619.

54 Leland G. Neuberg, "What Can Social Policy Analysts and Planners Learn from Social Experiments," *APA Journal*, Winter 1986, p. 68.

illustrative of experiments whose interpretation has been controversial. Nonetheless they may have influenced public policy based on initial reporting which was called into question later.

Income Maintenance. The Seattle and Denver income maintenance experiment (SIME/DIME) was the most ambitious effort devoted to the study of income support and has arguably influenced social policy.[55] The families involved received a guaranteed income equal to 95, 120, or 140 percent of the poverty line, and a variety of tax rates were applied to the participants' earned income for each subsample. This carefully designed experiment involved 4,790 families at a cost of $195 million (1991 dollars). SIME/DIME was the most comprehensive of four experiments undertaken between 1967 and 1982 to test the impact of negative income tax programs.[56] The SIME/DIME analysts concluded that the negative income tax tested in the experiment increased the rate of marital dissolution by 40 to 60 percent. Their findings probably influenced the 1978 congressional welfare reform deliberations, and they continue to be cited as significant and reliable evidence of important effects of guaranteed income programs. Senator Daniel Patrick Moynihan, chairman of the U.S. Senate Finance Committee's Subcommittee on Public Assistance, expressed the dominant interpretation of the experimental findings "... were we wrong about a guaranteed income! Seemingly it is calamitous. It increases family dissolution by some 70 percent, decreases work, etc. Such is now the state of the science, and it seems to me that we are honor bound to abide by it for the moment."[57]

---

55 Michael T. Hannan, Nancy Brandon Tuma, and Lyle P. Groevenveld, "Income and Marital Events: Evidence from an Income-Maintenance Experiment," *American Journal of Sociology*, May 1977, pp. 1186-1211; Leland Gerson Neuberg, "What Can Social Policy Analysts and Planners Learn from Social Experiments?," *Journal of the American Planning Association*, Winter 1986, pp. 68-70, 73.

56 David Greenberg and Mark Shroder, *Digest of Social Experiments* (Madison, WI: Institute for Research on Poverty, Special Report #52, 1991), p. 11.

57 Letter to William F. Buckley, Jr. "Notes and Asides," *National Review*, September 1978, p. 1196.

Other analysts have questioned the validity of the initial interpretation of the SIME/DIME marital stability data.[58] The critics have concluded that, contrary to the earlier analysis, the income maintenance plans tested in SIME/DIME had no effect on the marital stability of enrollees. They obtained their conclusions by isolating the effects of the income maintenance plans from the effects of training components included in some of the experimental treatments, utilizing a larger set of subsamples than the original analysts used. They also employed different adjustments for attrition and marital reconciliation, and sorted out the temporary and permanent effects of the treatment in a different manner. The revised data based on the disaggregated analysis yielded no conclusive evidence about the impact of SIME/DIME findings on family stability. Other critics of the experiment argued that in order to obtain reliable predictions, random experiments should be based on sufficient probabilistic samples and provide for a wider range of treatments than tested by SIME/DIME. These critics offered neither any estimate about the costs the proposed experiments would entail, nor did they indicate who would fund such experiments. The controversy persists as neither the original SIME/DIME analysts nor their critics have altered their positions in any significant way.[59] By the time the controversy surfaced, public policy interest in the negative income tax had waned, and, considering the cost of the experiment, funds for similar experiments were not available.

Even if the analysts had reached a consensus about the impact of negative income tax upon marital stability, the relevance of SIME/DIME for public policy would still remain

---

58 Glen G. Cain and Douglas A. Wissoker, "Do Income Maintenance Programs Break up Marriages? A revaluation of SIME/DIME," *Focus* (Madison, WI: Institute for Research on Poverty, Winter 1987).

59 *Lessons from the Income Maintenance Experiments: Proceedings of a Conference Held in September 1986*, Alice H. Munnell, ed, Conference Series no. 30 (Boston: Federal Reserve Bank of Boston, 1987); Spencer Rich, "Study Challenges Theory on Family Breakups: Few Husband-Wife Splits Attributed to Guaranteed-Income Plans," *Washington Post*, March 29, 1988, p. A15; Richard P. Nathan, *Social Science in Government* (New York: Basic Books, 1988), pp. 57-60.

problematic. For the purpose of the experiment, legal marital status was not a condition for receiving income support.[60] It is highly doubtful whether policymakers would adopt the SIME/DIME definition of family. Therefore, in case a negative income tax were ever adopted the findings of the experiment would be of limited practical use.

Criminal Justice. In 1962 Congress mandated federally-funded training to prepare prisoners for employment upon release. A 1972 review concluded that the training efforts had no impact upon recidivism.[61] The U.S. Department of Labor sponsored two randomized field experiments conducted in 1972-1974 and 1976-1977: the Living Insurance for Ex-Offenders (LIFE) experiment and the Transitional Aid Research Project.[62] These experiments were intended to determine whether 13 to 21 weeks of transitional financial aid, equivalent to average state unemployment insurance payments, and other services would reduce recidivism and enhance job placement and earnings of released prisoners. The LIFE experiment was conducted in Baltimore, Maryland, at a cost of $705,000 (1991 dollars) with a sample of 432 released prisoners. The analysts of the LIFE experiment concluded that income support reduced the recidivism rate of the treatment group by 15 percent and increased employment as well as enrollment in training or schooling. The more ambitious TARP experiment followed. It involved a cost of $8.3 million (1991 dollars) and nearly 3800 persons released from Texas and Georgia state prisons.[63] The investigators concluded that TARP payments (provided under different administrative rules than the LIFE experiment) produced no reduction in recidivism. Clearly, for the purposes of policy

60 Felicity Skidmore, "Overview of the Seattle-Denver Income Maintenance Experiment Final Report" in *Evaluation Studies, Review Annual*, Vol. 10, Linda A. Aiken and Barbara H. Kehrer, eds (Beverly Hills, CA, 1985), pp. 321-325.

61 Robert Taggart, *The Prison of Unemployment* (Baltimore: The Johns Hopkins University Press, 1972), pp. 96-97.

62 Richard A. Burk, Robert F. Boruch, David L. Chambers, Peter H. Rossi, and Anne D. White, "Social Policy Experimentation, A Policy Paper," *Evaluation Review*, August 1985, pp. 399-403.

63 David Greenberg and Mark Shroeder, *Digest of Social Experiments* (Madison, WI: Institute for Research on Poverty, IRP Special Report No. 52, May 1991) pp. 171-75.

analysis, it was important to understand the conflicting outcomes of the two projects. The TARP researchers concluded that the observed outcomes reflected the counteracting effects of work-disincentives and income on the rate of recidivism. Relying in part on collateral nonexperimental evidence of a negative association between hours worked and recidivism, they produced estimates of the magnitude of these effects. The estimates suggested that controlling for the work-disincentive effects associated with the particular features of the TARP design, income decreased the rearrest rate in the year following release.

A TARP advisory committee member expressed skepticism about the significance of the results, concluding that the experimental evidence was inadequate to distinguish between the TARP evaluators' assumptions and alternative specifications which might have led to different inferences.[64] Although he praised the quality of the experiment, he stated that "scholars disagree sharply on its evaluation. This seems to me an intolerable situation because it invites general contempt for social science research..."[65]

The participants in this acrimonious controversy have not resolved whether the results are of any use in the design of future experiments. Equally, they have reached no agreement on the underlying question of the proper use of behavioral modelling in the interpretation of future experimental results. Instead, they agreed only that more experiments are necessary. The question remains whether the experiments justified the costs or whether the experimenters could have obtained the desired results by some alternative method at less cost.

The Claims of Experimentalists

Proponents have contended that well-funded social experiments will redeem the long-standing promises of policy analysts to provide policymakers with a secure factual basis for rational consensus about the effects of social programs.

64 Hans Zeisel, "Disagreement Over the Evaluation of a Controlled Experiment," Peter H. Rossi, Richard A. Berk, and Kenneth J. Lenihan, "Saying It Wrong With Figures: A Comment on Zeisel," Zeisel, "Hans Zeisel Concludes the Debate," *American Journal of Sociology*, September 1982, pp. 378-396.
65 *Ibid.*, p. 395.

Aside from the high costs involved, are these expectations justified? The contributions of the experimentalists undeniably have been significant. They build cumulatively upon the advances of modern statistical theory and on over two decades of experience with large-scale social policy experiments. Social experiments have attracted the sustained attention of many noted social policy analysts, directly and indirectly produced many important methodological innovations of wide importance, and arguably influenced the course of social policy. Using their considerable resources, the experimentalists' widely-publicized research findings call into question the reliability of econometric evaluation techniques and other evaluation methodologies.[66] Detractors charge that the prominence of the experimentalists in the social evaluations arena may have been due to the extraordinary power of their rhetoric to evoke simple, widely-held images of the way science works.[67] Emphasis on experimentation, according to one critic, "is a natural consequence of the currently fashionable - but factually and intellectually unsupported - belief in social experimentation as *the* method of choice in program evaluation."[68]

Granting the contributions of the experimentalists, evidence and experience suggest that the expectations they have raised will prove incapable of fulfillment. Critics have charged that the results achieved by social experiments as well as other evaluation techniques are frequently ambiguous,

---

66 Robert F. Boruch, "Comparative Aspects of Randomized Experiments for Planning and Evaluation," in *Social Science and Government: Comparative Essays on Britain and the United States*, Martin Bulmer, ed (Cambridge: Cambridge University Press, 1987), p. 329.

67 Geoffrey Cantor, "The Rhetoric of Experiment," in David Gooding, Trevor Pinch, and Simon Schaffer, *The Uses of Experiment: Studies in the Natural Sciences* (Cambridge: Cambridge University Press, 1989), pp. 159-180.

68 James J. Heckman, "Basic Knowledge - Not Black Box Evaluations," *Focus* (Madison, WI: Institute for Research on Poverty, March 1992), p. 24; James J. Heckman and Joseph Hotz, "Are Classical Experiments Necessary for Evaluating the Impact of Manpower Training Programs? A Critical Evaluation," *Proceedings of the Fortieth Annual Meeting*, Industrial Relations Research Association, Barbara D. Dennis, ed (Madison, WI, The Association, 1987), p. 293.

and sometimes misleading, guides to policy.[69] The prospect that new experiments will resolve the major sources of this ambiguity is remote, and this effort is likely to become "a misplaced search for certitude."[70] Equally important, the skeptical evaluation of nonexperimental methods by random assignment proponents already has begun to appear anachronistic. The limitations that the critics have identified were most apparent in the early developments of these methods that came into wide use a generation ago; but nonexperimental methods and the sophistication with which analysts have employed them have evolved no less than experimental methods in the intervening period. As methods for drawing causal inferences from nonexperimental data continue to evolve, the stability of the experimentalist consensus can be expected to be called into question.

The methodological developments on which the proponents of social experiments have focused attention may still augur progress if the researchers would moderate the claims that the experimentalist program is likely to achieve. Along such a course, experimentation would constitute an important but not paramount element of the evaluation process, to be developed in concert with appropriate theory, taking cognizance of nonexperimental, microsimulation, qualitative, and ethnographic methods.[71] This is in stark contrast to the interests of experimentalists. The high cost of experiments encourages research funders in and outside the government to beggar other elements of potentially useful research. The concern is that, as long as the stronger versions of experimentalist rhetoric prevail and receive credence, researchers will follow the rewards paid in the coin of the realm.

---

69 James J. Heckman, "Randomization and Social Policy Evaluation," in *Evaluating Welfare and Training Programs*, Charles F. Manski and Irwin Garfinkel, eds, (Boston, MA: Harvard University Press, 1992), pp. 202-4.

70 Barry Bozman and Jane Massey, "Investing in Policy Evaluation," *Public Administration Review*, May/June 1982, p. 257.

71 Eleanor Chelimsky, "Old Patterns and New Directions in Program Evaluation," in *Program Evaluation: Patterns and Directions*, Eleanor Chelimsky, ed, *Public Administration Review* (1985), p. 14; Carrol Hirschon Weiss, "Evaluating Social Programs, What Have We Learned," *Society* November/December 1987, p. 43.

Qualitative Evaluation

Although quantitative methods have been the preferred approach to program evaluation during the past quarter century, there is a long history of dispute about the wisdom of that practice. Indeed, qualitative analysis preceded quantitative in significant usage, and the dispute didn't develop until quantitative methods began to dominate in the late 1960s. While proponents of quantitative approaches extolled the virtues of narrowly defining research questions and seeking answers through "objective" data gathering and analysis, critics maintained that such studies were often too narrowly focused to be relevant, and called for research to be more participant-focused.[72]

The contention over methodology unquestionably reflected inter-disciplinary turf battles, with economists generally advocating quantitative approaches, and sociologists, anthropologists, and political scientists preferring qualitative methods. Sometimes the controversy focuses on whether nonquantitative methods might have been used to obtain the desired information. The real issue, however, is how both methods can be used to the greatest advantage by stressing their separate strengths, rather than by dwelling on the other method's weaknesses.

Qualitative studies tend to gather information or "data" on a limited number of cases, but in much greater depth and detail than quantitative approaches, and with more emphasis on process than outcome. A strength of qualitative studies is their interest in exploratory findings, with far fewer preconceived notions or hypotheses to be tested than quantitative evaluations; data are often gathered directly from participants, through interviews or observations, and are apt to be discovery-oriented, free in form or open-ended as well as descriptive and contextual. Qualitative evaluation is generally ethnographic in nature, focusing on in-depth reactions of program participants. The approach is intrusive as far as the participants are concerned, probably causing a Hawthorne effect, but without attempting to impose any changes on the structure of the program. In contrast, quantitative researchers typically are impersonal. When data are gathered, experiments impose

---

72 Robert S. Weiss and Martin Rein, "The Evaluation of Broad-Aim Programs: Experimental Design, Its Difficulties, and an Alternative," *Administrative Science Quarterly*, March 1970, pp. 97-109.

constraints on programs (e.g., random assignment) to assure that the data will fit into pre-determined, standardized categories that are set up to test specific hypotheses that, for practical reasons, must be limited in number.[73]

Two key issues with respect to qualitative evaluations are the extent of their replication and generalization. Critics suggest that a weakness of qualitative studies is that different analysts doing the same research on the same sites under the same conditions might react differently, record different data, and reach different conclusions.[74] Quantitative studies employing identical analysis on different sample data might also get different results. Lacking rigorous rules and methodology, the qualitative analyst has greater freedom than the quantitative counterpart in structuring an evaluative study. Quantitative studies are presumed to be largely replicable, so that given the same data and performing comparable analysis, different researchers would reach the same conclusions, making the study reproducible. This criticism assumes that qualitative analysts ignore accepted social science research practices relating to reliability, accuracy, and objectivity. In fact, neither quantitative nor qualitative analysts have a monopoly on "sloppy" evaluations.

A Brookings field evaluation of job creation programs illustrates the gains of flexibility and insight obtained by linking qualitative with quantitative approaches. The evaluators, viewing the program from the perspective of recipient local governments, concluded that job creation was substantially greater than suggested by estimates derived from quantitative studies. An econometric study produced by the U.S. Department of Labor had concluded that the job displacement of federally funded job creation programs might exceed 50

---

[73] Michael Quinn Patton, *How to Use Qualitative Methods in Evaluation* (Beverly Hills, CA: SAGE Publications, 1987) in David D. Williams, ed, *Naturalistic Evaluation* (San Francisco: Jossey-Bass Inc., Publishers, 1986).

[74] S. E. Runyan, cited in Louis H. Kidder and Michelle Fine, "Qualitative and Quantitative Methods: When Stories Converge," in Melvin M. Mark and R. Lance Shotland, eds, *Multiple Methods in Program Evaluation* (San Francisco: Jossey Bass Inc., 1987), pp. 57-75.

percent.[75] That quantitative approach essentially used aggregated data to compare government employment before and after the introduction of public service employment (PSE) and netted out PSE slots awarded to determine the amount of job creation. Suppose a governmental unit had 1000 public jobs in year A, a number projected to remain steady in year B. If the unit was allocated 100 PSE slots, and public employment in year B was 1050, econometricians estimated job creation to be 50 percent, with substitution also 50 percent. The Brookings researchers, however, considered a PSE slot to represent job creation even if it filled a position that was formerly funded by the local government, but would not have been maintained without PSE dollars (which they defined as "program maintenance"). In the above example, they may have discovered that local economic stress would have lowered public employment to 970 in the absence of PSE, so that job creation was really 80 percent. When appropriate micro-level data were used to quantitatively estimate job creation, the estimates coincided with those made using the qualitative approach.[76]

Quantitative researchers frequently fail to consider the institutional environments in which programs operate.[77] In 1980, one-third of the slots allocated to local authorities were passed through to non-profit agencies. In addition, the rate of pass-through was much higher for service delivery areas in relatively sound fiscal shape, while it was much lower in fiscally distressed areas.[78] A quantitative study looking only at total public employment in a recipient jurisdiction

---

[75] George E. Johnson, "Evaluation Questions for the Comprehensive Employment and Training Act of 1973" (Washington: U.S. Department of Labor, Office of Assistant Secretary for Policy, Evaluation, and Research, July 1974).

[76] Charles F. Adams, Jr., Robert F. Cook, and Arthur J. Maurice, "A Pooled Time-Series Analysis of the Job Creation Impact of Public Service Employment Grants to Large Cities," *Journal of Human Resources*, Spring 1983, pp. 283-294.

[77] Richard P. Nathan, *Social-Science in Government* (New York: Basic Books, 1988), pp. 178-81.

[78] Robert F. Cook, Charles F. Adams, Jr., V. Lane Rawlins, and Associates, *Public Service Employment: The Experience of a Decade* (Kalamazoo, MI: The W.E. Upjohn Institute for Employment Research, 1985), pp. 41-42.

would thus greatly over-estimate displacement without adjusting for the number of slots allocated to non-profit agencies.

A more traditional example of ethnographic research would focus on the program participants themselves. A leading qualitative analyst offered the following sample questions as a guide for evaluating an employment and training program:

> "What has the trainee done in the program: activities? interactions? products? work performed? What are the trainee's current work skills? What things can the trainee do that are marketable? How has the trainee been affected by the program in areas other than job skills—feelings about self? attitudes toward work? aspirations? interpersonal skills? spin-offs? What are the trainee's plans for the future—work plans? income expectations? lifestyle expectations/plans? What does the trainee think of the program—strengths? weaknesses? things liked? things disliked? best components? poor components? things that should be changed?"[79]

The guide is not comprehensive, and the interviewer would be expected to explore additional areas of interest indicated by the respondents' replies. The guide itself would likely be the product of several preliminary interviews in search of questions, rather than answers. Further, the interviewer might conduct multiple interviews with the same respondents over an extended time period, and would also spend time observing the respondents within the natural setting of the program. Skeptics argue that the information might be of interest, but would yield little about the impact of the program on employment and earnings.

It is instructive to compare these questions with the approach a quantitative study might take. Both approaches might seek similar data: what did a participant do in a training program? A quantitative study might measure the participants activities in terms of time spent in specific training programs — job search assistance, classroom training, on-the-job train-

79 Michael Quinn Patton, *Qualitative Evaluation Methods* (Beverly Hills, CA: SAGE Publications, Inc., 1980), p. 201, citing John Lofland, *Analyzing Social Settings* (Belmont, CA: Wadsworth, 1971).

ing. The findings would be easily aggregated for statistical analysis, but it could also mask the reactions of individuals to identical treatments and the effect of the training they received. If two people took the same classroom training, one might have found it helpful, while the other repetitive of prior experiences and useless. It is not far fetched to assume that in carrying out their analysis quantitative evaluators are not concerned about what happens to individuals - a number is the end all of their efforts and the individual is ignored in the process. Quantitative analysts might well ponder the conclusion of one critic "...the measurable is not necessarily the important."[80]

Differences between the two approaches are apparent. The quantitative analysts, for example, would identify pre-treatment characteristics, such as education, work experience, prior training, pre-program earnings over some period. In an experiment random assignment would be assumed to eliminate any unobserved differences between participants receiving treatment and those assigned to a control group. Qualitative analysts, on the other hand, would identify what participants, based on their own reporting, bring to the program (and the researcher's assessment of that reporting). The participants' self-image, work attitudes, and aspirations could be supplemented and become research questions in the qualitative study, at least partly as outcome variables, instead of being assumed away up front on the basis of random selection, and ignored as outcome variables. While a quantitative analyst might recommend changes in the program based on measures of program impact, the qualitative researcher relies more on the participants' perspectives that would undoubtedly include issues never considered by quantitative analysts.

Qualitative analysts reject both "objectivity" claims in quantitative studies and "subjectivity" in qualitative studies. Objectivity suggests that researchers are seeking a single correct answer, or *truth*, while qualitative researchers acknowledge multiple "truths" from varied perspectives. The issue, instead, should be framed in terms of neutrality, fairness, and balance. There should be no more reason to suspect the conclusions of a well done qualitative than quantitative

80 Ida R. Hoos, *Systems Analysis in Public Policy* (Berkley, CA: University of California Press, 1972), p. 181.

study.[81] In fact, however, experimentalists have tended to neglect qualitative evaluations. Although experimentalists have failed to deliver on their promises, they won out because they always come up with a hard number, not necessarily stating or recognizing the assumptions they made to arrive at the number.[82]

Sound qualitative studies preclude generalizations not only because they are normally based on small samples, but also because qualitative analysts tend to view generalizations as unlikely to hold up well over time and across areas. However, recognizing the potential usefulness of generalization, some qualitative practitioners offered suggestions for making their findings susceptible, to reasonable "extrapolation." Such extrapolation, though perhaps less "rigorous" than the product of experimental analysis, may be just as useful for generalization if not more so.[83]

The Evaluation Industry

The debates among the practitioners of different evaluation methods have focused on substantive issues, but institutional factors have also played a role in the ongoing controversies. While the thriving evaluation literature has focused on substantive findings and methodological issues, the practitioners have devoted little public attention to the structure of their trade and the institutions in which they operate. The limited attention evaluators pay to the structure of their trade is not necessarily due to excessive modesty. It is more reasonable to assume that the neglect is due to methodological factors. The institutional impacts on evaluation are virtually impossible to capture in quantitative terms. Any unflattering attempt to tackle the institutional issues is almost certain to be disputed as biased misinformation — a safe

81 Michael Quinn Patton, *How to Use Qualitative Methods Evaluation* (Beverly Hills, CA: SAGE Publications, 1987), pp. 166-67.

82 Carol Hirschon Weiss, "Evaluating Social Programs: What Have We Learned?" *Society*, November/December 1987, p. 43.

83 Michael Quinn Patton, *How to Use Qualitative Methods in Evaluation* (Beverly Hills, CA: SAGE Publications, 1987), p. 168.

accusation since any observations are necessarily subjective and difficult to verify with objective data. These conditions discourage critical examination of institutional issues. However, the practical consideration of accounting for the neglect of institutional aspects relating to evaluation should not be ignored because the institutional arrangements for evaluations frequently drive substantive decisions and the methodologies used. The fact, whether or not it is acknowledged, is that institutional structures of the evaluation industry influence the federal government's social program evaluation policies. The purpose that evaluation serves, the way it is done, and the manner in which evaluation findings are incorporated into policy are directly affected by the way the evaluation is structured.

Executive Agencies

While most evaluation research is done by individual analysts and organizations outside the government, the federal government is overwhelmingly the largest single funder of evaluation studies. But the leading government evaluation agency, the General Accounting Office (GAO), suggests that all is not well within the federal evaluation community. According to a 1988 GAO report, both program evaluation capability by federal executive agencies and the data collection needed for sound evaluation were "seriously eroded" during the Reagan years. The decline in evaluation efforts by federal non-defense executive agencies in the 1980s reversed a rapid growth during the 1970s. Although precise numbers are difficult to obtain, a range of estimates suggests that annual federal spending by executive agencies for evaluations of non-defense programs grew tenfold between 1969 and 1980, reaching over $300 million (in 1991 dollars).[84] Adjusted for inflation, funding for evaluation, according to one estimate, fell 45 percent between 1980 and 1988. Professional staff in federal agency evaluation units were cut 22 percent between 1980 and 1984 (from about 1,500 to 1,200), while a subset of 15 major agencies had lost 52 percent of their professional staff between 1980 and 1988. The number of evaluations pro-

[84] William R. Shadish, Jr., Thomas D. Cook, and Laura C. Leviton, *Foundations of Program Evaluation: Theories of Practice* (Newbury Park, CA: Sage Publications, Inc., 1991), p. 27; Eleanor Chelimsky, David Cordray, and Lois-ellin Datta, "Federal Evaluation: The Pendulum Has Swung Too Far," *Evaluation Forum*, Fall 1989, pp. 92-94.

duced by executive agencies dropped only 3 percent between 1980 and 1984, but studies tended to be less complex. Unless they are mandated by Congress, the studies also were increasingly likely to be for internal use only. The decline in federal agencies evaluation personnel has forced the agencies to contract more evaluations with private firms. Reduced data collection, meanwhile, made it increasingly problematic that outside organizations could fill the gap left by the reduction in executive agency evaluation.[85] The Bush administration has indicated an interest in correcting the situation, but results have been meager.[86]

Staff shortages may place constraints on the quality of evaluation contracts. The sharp cuts in evaluation personnel has most likely adversely affected the ability of staff to develop requests for proposals, review bids, monitor contracts, or assess completed evaluations. Indeed, in some cases agency personnel have had to rely upon outside contractors or consultants to prepare the requests for proposals. The GAO concern about the state of evaluation activities of social programs in the executive agencies seems to be well justified.

Congressional Agencies

In the 1970s Congress established independent policy analysis as part of the legislative branch. To achieve this goal, Congress expanded the scope and responsibilities of two existing agencies - the Congressional Research Service and the General Accounting Office - and established two new ones — the Office of Technology Assessment and the Congressional Budget Office. These agencies have replaced executive agencies as the prime source for the analysis and evaluation of federal social programs.

Evaluation takes on a number of forms in the legislative branch. Congressional committees' work ranges from

---

85 Comptroller General of the U.S., *Program Evaluation Issues*, Transition Series Report to Congress, OCG-89-8-TR (Washington: U.S. General Accounting Office, November 1988). — *Federal Evaluation: Fewer Units, Reduced Resources, Different Studies From 1980*, Report to Congress, PEMD-87-9 (Washington: U.S. General Accounting Office, January 1987).

86 Economic Report of the President (Washington: Government Printing Office, February 1992), pp. 245, 278.

anecdotal assessments to more careful and systematic reviews. The General Accounting Office primarily conducts original reviews of programs, while the Congressional Research Service focuses on the analysis of evaluation literature prepared by the executive branch and others to produce a synthesis of current views. It has also prepared original studies dealing with program participation and other subjects. The Congressional Budget Office focuses its activities on estimating the costs of proposed federal programs. The binding force among all the legislative branch's evaluation work is its constitutional and institutional oversight role. The legislative branch has the duty to keep tabs on how the executive branch is carrying out congressional mandates. But although it gives legitimacy to the congressional evaluation role, the constitutional basis does not necessarily guarantee the effectiveness of these efforts.

The estimated 1992 budget and personnel strength of the four support agencies follow:[87]

	Budget (in millions)	Personnel
Total	540.5	6,258
CBO	$22.9	238
CRS	$56.8	815
GAO	$440.0	5062
OTA	$20.8	143

Under the 1970 Legislative Reorganization Act, which authorized research and policy analysis activities by the agency, the Legislative Reference Service was renamed the Congressional Research Service (CRS), and increased funding followed. Within less than a decade, CRS staff positions had more than doubled to over 800.[88] CRS officials maintain that the agency does not perform program evaluations, as the term is normally used. By whatever name it is called, evaluation is a major function of CRS. Good policy analysis — the

87 *Budget of the United States Government, Fiscal Year 1993* (Washington: U.S. Government Printing Office, 1992), Appendix A.

88 Sar A. Levitan and Gregory K. Wurzburg, *Evaluating Federal Social Programs: An Uncertain Art* (Kalamazoo, MI: The W.E. Upjohn Institute for Employment Research, 1979), p. 76.

examination of alternative courses of action and of their implications — requires an assessment of prior experience. For that purpose, CRS makes much use of evaluations of executive agencies, the General Accounting Office, academic studies, and interest groups, but CRS also examines those evaluations, digests them, and incorporates the conclusions, if not the details, into policy analysis. In recent years CRS has expanded its functions by examining administration data and other data sources that evaluators may have neglected. For example, CRS analysts are major contributors to the annual "Green Book," the most comprehensive collection of information on federal social programs.[89]

The distinction between policy analysis and evaluation appears more semantic than substantive because the line of demarcation between the two is frequently blurred. Members of Congress have been known to pick the part of the analysis that suits their needs and views. A top official of CRS volunteered "...the legislative policy analyst must [not] behave as a political eunuch or ignore value premises completely." In fact "the legislative...analyst is free to range the entire spectrum of possibilities."[90]

In 1974, the Congressional Budget and Impoundment Control Act established the Congressional Budget Office (CBO). The agency's major function is to assist the appropriate congressional committees with estimating the impact of specific programs that affect the federal budget. As in the case of CRS, the CBO makes no formal recommendations. An agency publication states: "'On the one hand' and 'on the other' are phrases used frequently in CBO reports...." No doubt, users of CBO reports often choose which hand they prefer.[91]

---

89 U.S. Congress, House Committee on Ways and Means, *Green Book* (Washington: U.S. Government Printing Office, 1991, 1641 pages).

90 William H. Robinson, "Policy Analysis for Congress: Different Methods? Different Styles?" paper prepared for the Association for Public Policy Analysis and Management, October 18-20, 1990, San Francisco, CA.

91 Congressional Budget Office, "A Profile of the Congressional Budget," September 1990, p. 6.

The Office of Technology Assessment (OTA) dates back to 1972. Its mandate is to report to Congress on the application of technology "for achieving specific goals."[92] Despite its specialized responsibilities, the OTA has interpreted its jurisdiction broadly and has evaluated a number of social programs.

By far, the leading congressional evaluation agency is the General Accounting Office, with a budget four times that of CRS, CBO, and OTA combined. Established in 1921, the GAO largely limited its activities during its first half century to auditing functions, focusing almost exclusively on the legality of federal expenditures without passing judgment on the propriety of the outlays. It performed its first major evaluation when Congress mandated GAO to assess the operations of the Economic Opportunity Office, the Great Society antipoverty agency.[93] By default, the GAO now routinely performs much of the evaluation work that was formerly carried on by the executive agencies. The Legislative Reorganization Act of 1970 and the Congressional Budget and Impound Control Act of 1974 brought about the GAO's shift from its traditional audit function to leading federal evaluation agency. The 1970 Act required GAO to standardize budget and fiscal data in coordination with the Office of Management and Budget (OMB) and the Treasury Department. It also marked the beginning of regular GAO evaluations of program results. These new responsibilities were reinforced by the 1974 legislation, which authorized the establishment within GAO of an office for program evaluation. The Program Evaluation and Methodology Division (PEMD) is the leading group, but not the only division within GAO performing program evaluation; over 60 staff members have taken their expertise from PEMD to other divisions and regional offices.[94]

---

92 Carnegie Commission, *Science, Technology, and Congress* (New York: the Commission, October 1991), p. 23.

93 Harry S. Havens, "What We Were, Why We Are," *The G.A.O. Journal*, Winter/Spring 1990, p. 33; Sar A. Levitan, *The Great Society's Poor Law* (Baltimore, MD: The Johns Hopkins Press, 1969), pp. 310-311.

94 Eleanor Chelimsky, "Expanding G.A.O.'s Capabilities in Program Evaluation," *The G.A.O. Journal*, Winter/Spring 1990, pp. 43- 52.

GAO staffing has reflected its changing role. In 1972 only 0.2 percent of the agency's professional staff were social scientists; by 1978, that figure had risen to 6 percent. Hiring in the 1970s moved away from the accounting emphasis to public administration, operations research, engineering, statistics, and economics disciplines. Today's staff shows even greater diversity and is a marked departure from the historical dominance of accountants and bookkeepers, who are now hired almost exclusively to do the agency's accounting and financial auditing work.

The effectiveness of GAO's evaluation of social programs hinges upon the methodological adequacy of its work and the relevance and utility of the findings. GAO's evaluation activities are driven by the reality that they must meet the pressing Congressional needs. Methodological correctness may at times have to be sacrificed to the imperatives of timeliness (which it does not always master) and to the need to pass judgments even when there is inadequate information for the formulation of sound evaluations.

The agencies that are the subject of GAO's investigations frequently attack GAO findings on methodological grounds. Although the criticisms are self-serving, they are not unfounded. For example, when the GAO contended that Maryland overstated the job creation achievements of its enterprise zone activities,[95] the state officials responded that the GAO assessment was based only on the experience of three rural areas, although 15 zones were in operation and the GAO failed to include the largest employer in one of the zones.[96]

Given the potential impact of any GAO evaluation, its findings should be qualified and recommendations should be made with caution where the methodological underpinnings or sampling procedures are weak. But this presents a

95 General Accounting Office "Enterprise Zones: Lessons from the Maryland Experience" (Washington: General Accounting Office, December 1988), pp. 3, 5, 42.

96 Sar A. Levitan and Elizabeth I. Miller, *Enterprise Zones: A Promise Based on Rhetoric* (Washington: The George Washington University, Center for Social Policy Studies, March 1992), pp. 37-38.

dilemma for the GAO. Required to respond with minimum delay to congressional requests, GAO must sometimes settle for evaluations based on weak methods and inadequate samples on the grounds that it is better than no review at all. If it is to remain valuable to Congress, the GAO does not have the luxury of delaying its responses to congressional requests until it collects all the relevant data. The need for expeditious responses, however, does not exonerate GAO from stating explicitly the limits on inferences that may be drawn from its assessments. "[W]hat is most important sometimes," according to the top GAO evaluation official, "is not having the best design but having an adequate design that will bring the findings in at the time they were promised."[97] The GAO claims that its evaluations have become an integral part of congressional policymaking and have influenced the content and passage of some important legislation, including extending duration of medicare to mothers leaving AFDC, raising appropriations for assisting homeless youths, and many more.[98]

Virtually all GAO program evaluations include recommendations to program officials. Although the recommendations are aimed at clarifying congressional mandates and improving policy implementation, the GAO too frequently restates the provisions of the law establishing the program. It is not surprising, therefore, that following a path of least resistance, agencies usually concur with GAO recommendations, without changing their policies or operations. The agencies assume that once the report is filed, GAO will not return for a while and that Congress will fail to follow-up on the GAO recommendations. This is not to deny that the GAO reviews are sometimes especially perceptive and constructive.

Some congressional groups blame the GAO for the messages it delivers.[99] Claims that GAO has suffered a loss of professionalism and organizational integrity, and questions

---

97 Eleanor Chelimsky, "The Politics of Program Evaluation," *Society*, November/December 1987, p.31.

98 Eleanor Chelimsky, "Expanding G.A.O.'s Capabilities in Program Evaluation," *G.A.O. Journal*, Winter/Spring 1990, pp. 48-49.

99 Dana Priest, "GAO Analysts Often End Up in the Middle of Political Fray," *Washington Post*, April 29, 1992, p. A21.

about the efficiency of the organization, have subjected the agency to more public scrutiny than it is accustomed to receiving. GAO's response has largely been to deny the charges of bias in its work, to make greater efforts to work with its critics, and to welcome outside review of its operations.

The Grant and Contract Establishment

Much of the growth in evaluation undertaken by executive agencies is routinely contracted out to for-profit and non-profit consulting firms. A much smaller share of federal research dollars flows to academics, who have become much more entrepreneurial. Some scholars have affiliated with established organizations, although the payoff in the halls of academe in prestige and honor is for rigorous research which serves to advance a scholarly discipline rather than research that meets practical policy needs. The states have played a minor role in the evaluation arena. The ability to generalize to the national level from state-level evaluations is often limited.

As discussed earlier, the evaluation community outside the federal government has played an important role in shaping the nature and scope of program evaluation. The federal government may provide most of the funding, but it clearly should not, and does not, as a rule dictate the results of the evaluations. This fact raises questions about the players competing for the federal research dollar. A 1972 study suggested that the evaluation of federal social programs was a relatively competitive industry with few signs of dominance by a few large organizations.[100] This may not be true today, for although many small competitors remain an important segment of the evaluation industry, the largest organizations seem to be gaining an increasingly dominant position. Size itself provides organizations the opportunity to branch out of narrow specialties and become "all-purpose" providers of research services, capable of serving any agency on a wide variety of subjects. As organizations developed track records, it has become easier for them to get funding. This was not necessarily a clear case of the economically efficient driving out the inefficient or incompetent, however. The large organi-

100 Albert D. Biderman and Laurie M. Sharp, *The Competitive Evaluation Research Industry* (Washington: Bureau of Social Science Research, Inc., 1972).

zations have developed expertise in writing grant proposals, and have established personal relationships with the principal players providing the funding and those doing the research. This gives them significant advantages in the competition for funds where competition exists. The relationship has been reinforced by a free flow of intellectual talent between several organizations outside of the government and the agencies from which they typically receive funding.[101] Of the $59.5 million that the Employment and Training Administration of the U.S. Department of Labor expended during 1987-1990 for research and evaluation contracts, six organizations accounted for 42.7 percent of the total.[102] Similarly, six firms received 60 percent of the then Office of Education outlays for evaluation between 1975 and 1980.[103] In cases of sole-source contracting, minority set-asides, and "selected bidder listings," many individuals and organizations were effectively kept from competing at all.

Due to a lack of adequate professional staff, executive agencies are forced to rely increasingly on outside agencies to perform program evaluations. This approach contains potentially disturbing implications as it is apparently leading towards oligopoly-type structure of the evaluation industry. Whether these developments beget oligopoly-type pricing of evaluation services is uncertain, though it is debatable whether the federal government and foundations are getting their monies' worth for evaluation expenditures. Of far greater concern is a lack of competition over approach: if everyone competing for grants conforms to a limited set of "acceptable" ways of doing evaluation, the potential benefits resulting from letting a thousand flowers bloom are lost. Concentration of evaluation resources in a few organizations

101 Kenneth L. Kraemer, Seigfried Dickhoven, Susan Fallows Tierney, and John Leslie King, *Datawars: The Politics of Modeling in Federal Policymaking* (New York: Columbia University Press, 1987), pp. 90-97.

102 Raymond J. Uhalde, Administrator, Office of Strategic Planning and Development, Employment and Training Administration, letter dated October 15, 1991.

103 Peter H. Rossi and James R. Wright, "Evaluation Research," *Evaluation Studies Annual Review*, Vol. 11, David S. Cordrary and Mark W. Lipsey, eds (Newbury Park, CA: 1986), p. 51.

may discourage creativity and innovation as analysts and administrators follow the firm's "party line."[104] The seeming dominance of quantitative methods, and particularly randomized experiments, is one example of this development. Although debates will persist whether this is indeed the best available research approach to given questions, the fact remains that the existence of such a debate itself suggests a richness of detail that is being lost by allocating the bulk of evaluation dollars to a single approach.

Towards a More Balanced Evaluation Agenda

Advocates and practitioners of social policy experiments frequently claim that their method of estimating program impact offers policymakers better insights and more reliable knowledge about the effects of policy decisions than could be achieved by alternative methods. They regard the alternatives to experimentation, which involve statistical modeling of enrollees' and nonenrollees' program participation decisions or qualitative studies, with great skepticism.

Social experiments may involve significant outlays, are difficult to implement, produce sometimes doubtful results, and have their own characteristic sources of uncertainty, just as do nonexperimental studies. Experimentalists have established their dominance in program evaluation, glossing over the inherent shortcomings of their own approach. Although early nonexperimental methods sometimes produced highly uncertain estimates of program impact, the development of new, less assumption-dependent nonexperimental methods and advances in qualitative studies continue to be an area of vigorous research. It is misleading to view these methods only as an alternative to experimental methods. To the extent that the experimental process is more complex than the image of black-box science suggests, the experimental analysts frequently resort to these same methods in their own work.

104 Richard A Berk, Robert F. Boruch, David L. Chalmers, Peter H. Rossi, and Ann D. White, "Social Policy Experimentation: A Position Paper," *Evaluation Review*, August 1985, p. 416.

The advantages of the alternative (or complementary) approaches to social experiments in particular settings depend on diverse features of the institutional context of the programs under test as well as on the intrinsic features of the methods. Moreover, decisions about the appropriate approach to take in social policy analysis can never be taken in a costless environment. Nonexperimental evaluation can be used to test hypotheses that would require prohibitive costs if experimental studies were tried.[105] In determining the appropriateness of experimental methods in particular cases, research managers should balance cost against the diverse uncertainties of alternative methods.

The Youth Employment and Demonstration Projects Act (YEDPA) raises questions about the proper uses of experimental approaches and the misplaced attempts to use experimental methodology. Enacted in 1977, YEDPA was a combination of traditional work experience and skill training programs which attempted emphasis on experimental research projects. Nearly three years after YEDPA funding was terminated, the U.S. Department of Labor requested the National Research Council of the National Academy of Sciences to evaluate YEDPA operations.[106] The Council in turn appointed a fifteen person committee for the task, including mostly academics representing diverse disciplines — anthropology, economics, psychology, public administration, sociology, and statistics.

During its four years of operations, ending in 1981, YEDPA funded some 400 projects, nearly doubling annual federal outlays to $2 billion for youth employment and training programs. A record amount of $448 million (1991 dollars) was earmarked for designing and evaluating new approaches aimed at reducing joblessness among poor youths.

---

105 Burt S. Barnow, "The Uses and Limits of Social Experimentation," Industrial Relations Research Association, *Proceedings of the 40th Annual Meeting*, Barbara D. Dennis, ed (Madison, WI: The Association), p. 281.

106 Charles L. Betsey, Robinson H. Hollister, Jr., and Mary R. Papageorgiou, eds, *Youth Employment and Training Programs: The YEDPA Years* (Washington: The National Academy Press, 1985).

Clearly, the committee faced a challenging task. Congress required the Labor Department "to test the relative efficacy of different ways of dealing" with youth employment programs. To implement the congressional mandate the Labor Department had to act expeditiously, but as frequently happens when Congresss appropriates funds for new initiatives, it failed to authorize hiring the necessary personnel to do the job. Robert Taggart, a prominent analyst of social efforts in charge of the program, had to find outside help to mount the projects and to design plans for their evaluation.[107] Where outside help was unavailable, Taggart funded new organizations (intermediaries in the parlance of the trade) to plan, operate, and evaluate the projects entrusted to them.

Working under a rigorous deadline, it is not at all surprising that most of the hastily designed evaluations did not "meet rigorous scientific" criteria. It appears, however, that the council's committee started with a definite bias. "We believe," the committee stated, "that the feasibility of random assignment in program research [has] been demonstrated...."[108] In an appendix to the report, a committee member urged, "Randomized field experiments should be explicitly authorized in law and encouraged in evaluation policy...."[109] Given this orientation, the committee focused its assessments on 28 out of 400 projects. The committee chair defended the selection of acceptable project reports because they reflected "reasonable standards...if one were going to come to conclusions about program effectiveness." He added the committee found it "shocking" that it found only 28 usable reports.[110]

A reviewer of the council's report did not find the argument persuasive because the committee failed to evaluate the total YEDPA activities: "The report," he charged, "is less

---

107 *Youth Employment and Training Programs: The YEDPA Years*, Charles L. Betsey, Robinson G. Hollister, Jr., and Mary R. Papageorgion, eds (Washington: The National Academy Press, 1985), paper by Richard F. Elmore, pp. 299-303.

108 *Ibid.*, p. 30.

109 *Ibid.*, p. 231.

110 Robinson G. Hollister, Jr. "Youth Employment and Training Program," *Industrial and Labor Relations Review*, July 1987, p. 143.

about the accomplishment of YEDPA and more a commentary on the efficacy and utility of certain evaluation methodologies."[111] Indeed, it appears that the committee was easily shocked. The committee might have expressed surprise that given the resources, a small band of government "bureaucrats" succeeded in mounting in one year projects that significantly reduced youth unemployment. One experiment guaranteed work in 17 communities to poor youth aged 16 to 19 who had not finished school. The experiment increased the employment/population ratio of poor teens to 41 percent, compared with 25 percent in similar communities not participating in the experiment. The entitlement most dramatically influenced poor minority youth employment levels, raising them above the employment/population ratio of whites.[112] The evaluation committee apparently considered these "details" outside the scope of its mandate, although it devoted a section of its report to the project.

If anything, the National Research Council evaluation effort illustrates the futility of evaluating the achievements of social programs using rigorous experimentation methodology. It seems that the program operators and evaluators had different goals. The former were dedicated to creating jobs for unemployed youth while the committee was concerned with evaluation techniques. As a result our knowledge of YEDPA has been enriched very little.

---

111 Vernon M. Briggs, Jr., *Ibid.*, p. 138.

112 Judith M. Gueron, *Lessons From a Job Guarantee* (New York: Manpower Demonstration Research Corporation, June, 1984); Sar A. Levitan and Frank Gallo, *Spending to Save: Expanding Employment Opportunities* (Washington: Center for Social Policy Studies, The George Washington University, 1991), p. 17; Robert Taggart, testimony before U.S. Congress, House Subcommittee on Employment Opportunities, Committee on Education and Labor, *Hearings on Job Creation Proposals* (Washington: U.S. Government Printing Office, March 17, 1983), pp. 431, 437; *Youth Employment and Training Report*, pp. 151-158.

Efforts to develop formal frameworks that would guide research managers in this task have not produced any consensus.[113] Experimental methods are much more expensive than their nonexperimental alternatives, but proponents of experiments have not yet developed persuasive evidence that their favored techniques justify the extra costs as compared with alternative designs. A dedicated proponent of controlled experimentation concluded: "I think it is misguided to seek a global answer to the question: Are social experiments preferable to nonexperimental methods? The answer to this question must depend on the particular hypothesis being examined."[114]

When the Employment and Training Administration of the U.S. Department of Labor embarked on an evaluation of the Job Training Partnership Act, it appointed an advisory panel to help design the evaluation. The panel unanimously recommended an experimental approach. The members of the committee recognized, however, that it would be too costly to collect a sufficiently large sample to measure the various treatments offered by JTPA including classroom training, on the job training, and job search, and the varied environments in which the program operated. The panel therefore recommended, according to its chair, that the evaluation "will have to rely on quasi-experimental methods and statistical and economic modeling to complete the program evaluation. Indeed, the two broad approaches are not substitutes for one another. Rather, they complement each other."[115] In addition to costs, substantive considerations also led the panel to recommend adding a quasi-experimental component to the evaluation design. The evaluation was to last several years, and nonrandom attrition was bound to occur which would cast doubts

---

113 Gary Burtless and Larry L. Orr, "Are Classical Experiments Needed for Manpower Policy?," *Journal of Human Resources*, Fall 1986, pp. 606-639.

114 Gary Burtless, "The Social and Scientific Value of Controlled Experimentation," in Industrial Relations Research Association, *Proceedings of the 40th Annual Meeting*, Barbara D. Dennis, ed (Madison, WI: The Association), p. 269.

115 Ernst W. Stromsdorfer, "Evaluating the Net Impact of Employment and Training Programs: Some Lessons in Methodology," *Evaluation Forum*, February 1991, p. 61.

about the reliability of the findings. The Labor Department accepted the recommendation of its advisory panel, and, as noted earlier, the evaluation is still continuing six years after its initiation. The department scaled back all other JTPA research and dropped a longitudinal study of JTPA enrollees. Consequently, most of the available information about the functioning of the program in the last five years has been produced by the General Accounting Office and not the Department of Labor. By putting a major share of its research eggs in one basket, the department has failed to undertake other studies which could have improved program operations.

GAO has also long recognized that relying exclusively on experimental designs fell short of discharging the agency's responsibilities. Experimental evaluations could provide (within the limitations discussed earlier) estimates of program impact. However, congressional users "always wanted to know why it happened and how the lessons learned might apply to a new program. So a quasi-experimental design had to be backed up not only by a comprehensive literature review, but an explanatory process evaluation or a set of case studies and an analysis of policy implications."[116]

Given the current state of our knowledge and resource constraints, random assignment experiments cannot be conducted in more than a small percentage of areas that interest policymakers. Each experiment uses resources that otherwise could fund the production of broadly applicable information bases that would enrich knowledge about program operations through sample surveys as well as enhance program services. There is a large and growing literature that suggests the need to integrate qualitative and quantitative methods.[117] Qualitative researchers have discovered the usefulness of computers for organizing, cataloging, storing, retrieving, and analyzing data.[118] Increasingly, social science

116 Eleanor Chelimsky, "Expanding G.A.O.'s Capabilities in Program Evaluation," *The G.A.O. Journal*, Winter/Spring 1990, p. 46.

117 Richard Nathan, *Social Science in Government: Uses and Misuses* (New York: Basic Books, 1988), pp. 200-201.

118 Harold G. Levine, "Principles of Data Storage and Retrieval for Use in Qualitative Evaluation," *Educational Evaluation and Policy Analysis* (Washington: American Educational Research Association, 1985), pp. 169-186.

researchers are questioning the wisdom of the predominant reliance on random assignment as the sole acceptable method for evaluating social programs and are developing econometric techniques for use in alternative methods. Their message has, however, been slow to reach the funders.

Both sides of the qualitative-quantitative debate would seem to be well-advised to employ multiple approaches to evaluation. To implement their approach, the state government agencies and foundations requesting and funding studies need to judiciously allocate the distribution of funds among competing methodologies. For a long time, the balance of funding has weighed heavily in favor of quantitative methods. Given the frequent weaknesses inherent in most such studies, and the contributions that qualitative studies appear capable of, it would seem that the time has come for that balance to shift, not necessarily to a different extreme, but at least toward more balanced shares of the research dollar.

Responsibility for a realistic appraisal of the uncertainty that is associated with social policy evaluation rests, of course, largely with the policy analysts themselves, for it is they who must change the way they work. Granted its shortcomings, greater emphasis on qualitative research and evaluation could provide better understanding of program operations, the aspirations and needs of participants, and how to improve program performance.

Ultimately, however, the most important force for change in the conduct of policy analysis must come from the major funders of evaluation research, both private foundations and government agencies. The evidence is lacking that the sponsors have paid sufficient consideration to the "bottom line." Have the outlays for the evaluation been commensurate with the returns, by whatever standards funders may establish? The latter could better achieve their objectives if they adopted measures making the award of contracts or grants, at a minimum, conditional on knowledge building and full disclosure of the limitations presented in the findings. In addition, each funder could design other conditions in line with the organization's goals. To the extent that evaluation proposals cannot accommodate the specified conditions, evaluation funders should consider the possibility of redirecting their resources to other areas such as the delivery and provision of services or in enriching data sources.

The Uncertain Returns

After a quarter century of unprecedented investment in evaluations of federal social programs, eminent social scientists as well as policymakers and funders of evaluations continue to question whether the investments have paid off.[119] Repeatedly, analysts have been unable to agree about the impact of social policy evaluation, and when disagreement has arisen, conflicting conclusions typically have been presented in ways that make the disagreement unintelligible and difficult to resolve. In addition to flaws in individual studies, the underlying institutional and cultural factors have contributed to the very limited progress achieved by evaluators of social programs.

As noted, quantitative analysts typically present the "bottom line" of their findings in a single number, rather than a range of estimates that reflect the inevitable degree of uncertainty in the findings. This pattern of practice ill serves both evaluation research and policy interests. It also fails to provide funders with information for rationally allocating resources among competing evaluation proposals and other activities.

Current practices are equally inadequate as a guide to choosing among social policy alternatives. Policymakers and administrators are consistently confronted with conflicting claims about the effects of their actions because they are not provided full descriptions of the sources of uncertainty associated with the estimates. Skepticism about the findings of policy evaluation research has various sources. Undeniably, the tendency of analysts to resort to obscurantism in producing and communicating their findings is a widespread problem. The findings are too frequently expressed in language comprehensible only by the initiated in the specialized scholarly fraternity (or sorority). Improved results can be attained by

---

119 Edward G. Manning, et al., "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment, *American Economic Review*, p. 272; James J. Heckman, "Social Science Research and Policy," *Journal of Human Resources*, Spring 1990, pp. 297-304; Harvey S. Rosen in Jerry A. Hausman and David A. Wise, eds, *Social Experimentation* (Chicago: University of Chicago Press, 1985), p. 4.

making more effective use of the various methodologies and by achieving a better balance in funding them. The recommended changes will not deliver the Holy Grail, but they might provide policymakers useful analysis which would lead to more effective social programs. Though we are not likely to ever achieve absolute precision or eliminate uncertainty, there is room for optimism that the evaluators' contributions will justify continued funding of their efforts.

THE AUTHOR

Sar A. Levitan is Research Professor of Economics and Director of The George Washington University Center for Social Policy Studies

Copies of EVALUATION OF FEDERAL SOCIAL PROGRAMS: AN UNCERTAIN IMPACT and other center papers may be obtained from Public Interest Publications, 3030 Clarendon Boulevard, Suite 200, Arlington, VA 22201, P.O. Box 229, Arlington, VA 22210, 1-800-537-9359, (703) 243-2252

Other Recent Papers Published by the Center:

ENTERPRISE ZONES: A PROMISE BASED ON RHETORIC by Sar A. Levitan and Elizabeth I. Miller

GOT TO LEARN TO EARN: PREPARING AMERICANS FOR WORK by Sar A. Levitan and Frank Gallo

SPENDING TO SAVE: EXPANDING EMPLOYMENT OPPORTUNITIES by Sar A. Levitan and Frank Gallo

THE PARADOX OF HOMELESSNESS IN AMERICA by Sar A. Levitan and Susan Schillmoeller

68

George Washington University
WASHINGTON DC

CENTER FOR
SOCIAL POLICY STUDIES
1717 K STREET, NW
SUITE 1200
WASHINGTON, DC 20006